DISCOVERY AND CHARACTERIZATION OF NOVEL RNA REPAIR SYSTEMS

BY

PEI WANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biochemistry
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Associate Professor Raven H. Huang, Chair
Professor Satish K. Nair
Professor Huimin Zhao
Assistant Professor Hong Jin

ABSTRACT

RNA is one of the major macromolecules essential for all known forms of life. RNA in the cell is subject to many types of damage as DNA, and the ubiquitous RNA degradation by surveillance machinery is an important way to respond RNA damage. However, RNA repair is an alternative way for cells to deal with RNA damage, which may play an important role in maintenance of cellular RNAs and even for cell survival.

RNA repair is the mechanism that rectifies the purposeful RNA damage during RNA processing or cellular stress. To counter the unexpected RNA breakage, RNA repair systems have evolved in some organisms to restore the normal structure and function of RNA. The first RNA repair system was found in *T4*phage, in which two proteins *T4*Pnkp and *T4*Rnl carried out RNA healing and sealing, respectively. The bacterial Pnkp/Hen1 complex has also been shown to repair ribotoxin-cleaved RNAs *in vitro*, but it was distinguished from *T4* systems by performing 3'-terminal 2'-O-methylation during RNA repair, which prevents the repaired RNA from repeated cleavage at the same site. Bacterial Pnkp and Hen1 appear pair-wise in the same operon in approximately 5% of known bacterial species. Although the bacterial Pnkp has been shown to possess kinase, phosphatase for RNA healing and all the signature motifs of a RNA ligase, it alone is not able to carry out RNA repair. In our study, we crystallized an active RNA ligase consisting of the C-terminal half of Pnkp (Pnkp-C) and the N-terminal half of Hen1 (Hen1-N) from *Clostridium thermocellum*, and provided the molecular basis for the ligase activation of bacterial Pnkp by Hen1. We also carried out a detailed functional study to pinpoint the activation step during RNA ligation. Guided by the sequence and structure, we created a series of point mutants for this new ligase and carried out biochemical assays. These studies provide additional insight into the mechanism of RNA ligation by Pnkp/Hen1.

Based on a comprehensive Blast search, a novel RNA repair system composed of three proteins (Pnkp1, Rnl and Hen1) has been found in 10 bacterial species. This new system possesses the features from both *T4*Pnkp/Rnl and bacterial Pnkp/Hen1 systems. Efficient *in vitro* RNA repair only occurred in the presence of all three proteins. We showed that these three proteins formed a heterohexamer *in vitro* that contains two copies of each active site (kinase, phosphatase, methyltransferase and ligase). We crystallized and solved the strucuture of the heterohexamer to gauge four different enzymatic activities. Based on the structural and biochemical studies, we propose a novel mechanism of processive RNA repair with efficient 2'-O-methylation. The study of bacterial Pnkp1/Rnl/Hen1 complex may shed light on the bacterial Pnkp/Hen1 system due to their similarities.

Both structural and biochemical assays indicated bacterial RNA repair systems have a broad range of RNA substrates, but we still have little knowledge of their biological functions *in vivo*. We therefore introduced genes encoding toxins and RNA repair proteins into bacteria through tightly controlled plasmids, and tested the inhibition and recovery of cellular growth upon the induction of gene expression. This method helped to identify potential toxin genes predicted by sequence alignment, and also facilitated the *in vivo* study by mimicking the environment the bacteria might be exposed to. If the RNA repair system provides surviving advantage for bacteria to defend themselves under stress, targeting this system may have the potential to limit the growth of pathogenic bacteria possessing the RNA repair systems.

In addition to bacterial RNA repair, I am also interested in important NTase fold protein, which constitutes a large and highly diverse superfamily of proteins. Almost all known members of NTase transfer NMP to a hydroxyl group of substrates including proteins, nucleic acids and small molecules. A newly identified subgroup of this superfamily, Mab21, is shown to be

essential for development or immune response. Particularly, Mab21D1 (cGAS) is important for interferon response activation by sensing dsDNA in cytosol and generating a second messenger molecule cGAMP. I crystallized and solved the structure of a close member Mab21D2, which is highly conserved in vertebrates but with unknown function. Through structural comparison to other NTase fold members as well as preliminary biochemical assays, we propose that it might work differently than cGAS.

ACKNOWLEDGMENTS

I would like to express my appreciation to the department of Biochemistry in University of Illinois Urbana-Champaign for giving me the opportunity and financial support to study here. I owe my deepest gratitude to my research advisor, Dr. Raven H. Huang. Without his guidance and persistent help, this dissertation would not have been possible. He has taught me a lot about how to examine the issues critically, how to make connections between pieces of knowledge and how to solve problems in an efficient and logical way. I am also very impressed by his enthusiasm and dedication for science, which helped me prepare for my future career. I am very grateful to all my labmates: Dr. Chiomui Chan, Kiruthika Selvadurai, Can Zhang, Amy Carruthers, David Christensen and Joe Seimetz for academic support and friendship. I would like to thank my committee members Dr. Satish K. Nair, Dr. Huimin Zhao, Dr. Hong Jin and Dr. Scott Silverman for their advice on my research. I am also thankful to Dr. James H. Morrissey, Dr. Susan K. Martinis, Dr. Linfeng Chen, Dr. John H. Cronan, Jeff Goldberg, Alejandra Stenger and Cindy Badger for their encouragement and help.

I have also made some excellent friends here: Dr. Zhi Li, Paween Mahinthichaichan, Weixin Tang, Yin Li, Hanchao Zhao, Yue Hao, Xinying Zong, Peiwen Wu, Chang Sun and Xinyun Cao, who were fundamental in supporting me during the most stressful and difficult moments.

I am particularly indebted to my family for supporting all my decisions, inspiring me whenever I need and loving me unconditionally all the time.

I would like to thank the National Science Foundation and National Institutes of Health for the research funding.

TABLE OF CONTENTS

CHAPTER 1: BACKGROUND

This background section covers the subjects of RNA damage, RNA repair and NTase superfamily.

1 RNA damage

DNA damage is a significant issue for all cells, since DNA serves as the biological information carrier in most organisms. Some agents, such as radiation or viral infection, can cause the DNA damage that leads to many kinds of diseases, particularly cancer where DNA repair commonly fails (1). Like DNA, RNA in the cell is also subject to damage. However, RNA damage and repair have not been a major focus of study due to ubiquitous RNA degradation by surveillance machinery (2).

1.1 Programmed RNA damage

Purposeful breakage in tRNA or mRNA is very common during RNA processing, such as splicing of intron-containing tRNAs and mRNAs (3-7). It may not be proper to classify normal RNA process as RNA damage, but the mechanism of splicing shares similarities with certain types of RNA damage caused by endogenous or environmental factors. Splicing occurs in all kingdoms or domains of life, but the types and extent are very different. In eukaryotes, most of the splicing events involve the protein-coding mRNA, which is mainly catalyzed by a RNA-protein complex known as the spliceosome (Figure1.1) (4). However, due to the lack of the spliceosome, splicing events are rare in prokaryotes, and most of them involve self-splicing of non-coding RNA. Spliceosomal and self-splicing usually involve two transesterification reactions: a particular 2'-OH group in RNA performs nucleophilic attack on the phosphodiester bond at a specific site and causes breakage of RNA; the released 3'-OH group at the broken end

can carry out the second transesterification reaction and ligate with the desired fragment of RNA. The splicing of tRNA is different from this two-step splicing event (5-7), in that the breakage is caused by endonucleases and the re-ligation is accomplished by ATP-dependent ligase.

1.2 RNA damage under cellular stress

Non-specific RNase might be enough for cytotoxicity. For example, Barnase (1,8), a ribonuclease produced by *Bacillus amyloliquefaciens*, is a potent ribonuclease and breaks down RNA indiscriminately. Barnase would be highly toxic to *E.coli* cells without co-expression of its inhibitor gene *barstar*.

For effective cell killing, it might be easier for cytotoxic RNase to attack specific RNA. Some toxins can target mRNA in order to cause RNA interference and translation inhibition. This type of RNA damage is usually caused by sequence-specific endoribonucleases. For example, MazF members can cleave mRNA at 3-, 5- or 7-base recognition sequences to control cell growth (1). The bacterial toxin RelE can recognize and cleave mRNA in the ribosomal A site, and it exhibits highest cleavage activity after the codon Py-Pu-G (9, 10).

Ribosomal RNAs are good targets for many kinds of proteinaceous toxins due to the conservation and importance of the ribosome. Ricin (plant origin) and α-sarcin (fungi origin) act on the sarcin-ricin loop (SRL loop) of the eukaryotic ribosome via either site-specific depurination or endonuclease activity (11- 13), while the newly found VapC20 from *Mycobacterium tuberculosis* can work efficiently on the SRL loop of bacterial ribosome (14). In addition, Colicin E3 acts effectively on the bacterial ribosome by specifically cleaving 16S rRNA at the 49^{th} phosphodiester bond from 3'end (15-17).

Like rRNAs, tRNAs constitute another conserved and essential RNA machinery for protein synthesis, but tRNAs are smaller and simpler with more divergent members (18-23). In response to T4 phage infection, *E.coli* prr strain could express PrrC that cleaves its own $tRNA^{Lys}$ between position 33 and 34. *E.coli* uses this suicidal strategy to limit phage infection (18). Colicin E5 and Colicin D can also serve as specific tRNases to stop protein synthesis (Figure1.2) (22): Colicin D cleaves all isoacceptor $tRNA^{Arg}$ (ICG, CCG, U*CU, CCU) between position 38 and 39; Colicin E5-sensitive tRNAs commonly contain a modified nucleotide, queuine, at the wobble site (position 34) of the anticodon (e.g.: $tRNA^{Tyr}$, $tRNA^{His}$, $tRNA^{Asp}$, $tRNA^{Asn}$), and the cleavage is usually between position 34 and 35. The aforementioned tRNases can prevent protein synthesis because the lack of specific tRNAs abolishes the incorporation of certain amino acids during translation.

1.3 The consequence of RNA damage

Programmed RNA damage is part of RNA processing, and RNA usually can restore its normal and functional form. During mRNA splicing, ligation of exons is usually accompanied with the release of an intron, which is catalyzed by the splicesome or by RNA itself.

Most toxins targeting RNA usually cause cell death either by depleting essential proteins (9,10) or destroying universal translation machinery (11-23). Toxins are usually generated to control cell growth: the majority is secreted out to kill other organisms, especially under harsh environmental condition; some toxins are generated to control growth in response to cellular stress such as virus infection (27).

Some anti-cancer chemotherapy agents can cause RNA damage, which is an important component of the drug action. The drugs perform functions similar to toxins, and kill the cancer

cells by causing the damage of essential RNAs. For example, the anti-cancer drug cisplatin can inhibit the translation in cancer cells by crosslinking mRNA and rRNA. Furthermore, similarly to DNA damage, RNA damage can also cause cell cycle arrest and induce apoptosis through some unknown mechanism (1).

2 RNA repair

It has been shown that cells transcribe more RNA than they accumulate, implying the existence of RNA degradation systems. RNA is degraded at the end of its useful life (2), which is long for rRNA, but short for excised introns. Damaged RNA can be rapidly degraded by the surveillance machinery through endonuclease or exonuclease activity. RNA repair is the alternative way for cells to deal with RNA damages, which may play an important role in maintenance of cellular RNAs (24). It has been shown that AlkB homologues not only reverse alkylation damage of DNA, but also repair methylated RNA (25). Moreover, the CCA-adding enzyme may have an important role in repairing the 3'-end of tRNA in *E.coli* (26). In our study, we focus on the breaking damage of RNA, in which the transesterification reaction carried out by toxins cleaves the RNA and generates 2', 3'-cyclic phosphate and 5'-OH termini (Figure 1.3).

2.1 *T4* RNA repair system

As mentioned above, *E.coli* prr strain can cleave its own $tRNA^{Lys}$ near the anticodon loop when infected by *T4* phage (18). However, the cleaved tRNA products are transient with wild type *T4* infection, but accumulate with pnk^- or rli^- *T4* infection (27). The genes *pnk* and *rli* in *T4* could encode proteins that reverse the tRNA damage in order to prevent the suicidal behavior of the host (Figure 1.3). *T4 pnk* encodes protein polynucleotide kinase-phosphatase (Pnkp), which

shows three enzymatic activities *in vitro*: its N-terminal domain belongs to the P-loop phosphotransferase family that can 1) phosphorylate the 5'-OH termini of the polynucleotide; its C-terminal domain belongs to the acylphosphatase family (specified by DxDxT motif), which can sequentially carry out 2) hydrolysis of 2', 3'-cyclic phosphadiester bond and 3) hydrolysis of 3'-terminal phosphomonoester bond. *T4*Pnkp can process both broken ends of RNAs and generate 5'-P and 3'-OH termini. *T4rli* encodes RNA ligase that can ligate the processed RNAs to restore it back to the original form (28, 29).

2.2 Bacterial RNA repair system Pnkp/Hen1

Our lab discovered that two bacterial proteins, Pnkp and Hen1, were also able to carry out RNA repair. Bacterial Pnkp is shown to have kinase, phosphatase and adenylyl transferase activities (30, 31), which are hallmarks of RNA repair. Bacterial Pnkp can heal the broken ends (32), but it alone cannot carry out repair. Within the same operon, there is another conserved gene encoding bacterial Hen1 (Hua enhancer) upstream of Pnkp (Figure 1.4) (33). These two proteins, Pnkp/Hen1, form a complex and repair RNA with the addition of a methyl group at the ligation junction (34). Due to the lack of 2'-OH at the cleavage site, the methylated repaired product is resistant to repeated damages. Hen1 in eukaryotes is a S-adenosylmethionine (AdoMet)-dependent RNA methyltransferase (MTase), which carries out 2'-O-methylation at the 3'-terminal nucleotide of RNA (35-37). It's also shown that methylation of small RNA is essential for its stability and physiological function (38- 40). Hen1 homologs found in plants, animals and bacteria have different sizes and show different domain arrangements. Bacterial Hen1 has the methyltransferase domain (MTase) at its C terminus, while its N-terminal domain is uncharacterized. Apart from the MTase domain, bacterial Hen1 has nothing in common with

Hen1 in eukaryotes (Figure 1.5) (33). Based on the previous studies, Pnkp/Hen1 is shown to have kinase, phosphatase, metyltransferase and ligase activities, and this system does not show obvious substrate specificity in terms of RNA sequence or structure: two broken ends that can be brought close enough have the potential to be repaired (41).

Approximately 5% of bacterial species with known genomic sequences possess the Pnkp/Hen1 RNA repair system (Figure 1.6). The distribution is uneven among branches of bacteria, and the majority of bacterial species possessing Pnkp/Hen1 belong to Actinobacteria. Particularly, within this branch, all Streptomyces species, known as the main source of antibiotic production, possess Pnkp/Hen1 system.

3. NTase

3.1 NTase superfamily

Nucleotidyltransferase (NTase) fold proteins constitute a large and highly diverse superfamily of proteins: almost all known members transfer nucleoside monophosphate (NMP) from nucleoside triphosphate (NTP) to the hydroxyl group of the acceptors such as proteins, nucleic acids or small molecules (Table 1.1) (42). Sequence analysis of the members reveals the conserved active site residues for this superfamily: hG[GS], [DE]h[DE]h and h[DE]h (h stands for a hydrophobic amino acid). Three conserved acidic residues are for the coordination of divalent ions and activation of acceptor hydroxyl group. The members of NTase superfamily are characterized by the common core structure containing 3 stranded β-sheet flanked by 4 α helices (Figure 1.7). This core structure is usually decorated by additional structural elements in different families.

NTase superfamily enzymes catalyze reactions involved in important biological

processes, such as DNA repair, RNA editing, antibiotic resistance, signal transduction and so on. For example, some DNA polymerases (e.g. Polβ, Polγ) play an important role in DNA repair by catalyzing DNA synthesis to fill the gap (43). KNTase (Kanamycin nucleotidyltransferase) catalyzes the NMP transfer to kanamycin and results in antibiotic resistance (44). Upon the stimulation of dsRNA, one NTase member OAS (2'-5'-oligoadenylate synthetase) can convert ATP to 2'-5'-linked oligoadenylate and activates RNase L in order to degrade viral or cellular RNA (45, 46).

Additional NTase fold superfamily members were identified through transitive meta-basic searches based on known NTase fold proteins. All the NTase fold proteins were assembled into 26 distict groups based on their sequence similarity (Table 1.2); the first 16 groups include known members while the remaining 10 are novel groups containing a majority of poorly characterized proteins (42). Some of the newly identified members may lack NTase function due to the lack of the essential catalytic residues.

3.2 Group XXIV of NTase: Mab21

Group XXIV is one of the newly identified NTase superfamily subgroups and the members possess the newly detected NTase domain followed by C-terminal PAP/OAS1 (PolyA polymerase and 2'5'-oligoadenylate synthetase) substrate binding domain (42, 47-50). This group is named after its small conserved eukaryotic Male-abnomal 21 domain, which was firstly reported in the nematode *Caenorhabditis elegans* and defined as a domain of around 360 amino acids (47). The protein encoded by one mab-21 gene (Mab21D6) is required for the choice of alternate cell fates in male tail.

Mab21 family has been studied as a highly divergent developmental protein family

(Danger), so the new members are usually denoted as D (with D standing for Danger, as in Mab21D6) (48). Danger sequences are differentially expressed in several tissues at various developmental stages, and they also exhibit species-specific expression levels. The study of the distribution and evolution of Danger genes in metazoa indicates that this protein family has continually been expanded during metazoan evolution and it may be responsible for the emergence of different cell types (Figure 1.8A. The names of proteins are changed according to the new study for the rest of this thesis, and the comparison is shown in Figure 1.8B). The members in this family are very divergent, and the sequence comparison of human Mab21 proteins shows that they can be correlated when the cutoff value is as high as 1e-5 (Figure 1.8C).

3.3 Mab21D1 (cGAS)

Recently, one member of the Mab21 family, Mab21D1 (also called cGAS: cyclic GAMP synthetase) was shown to participate in immune response upon the stimulation of cytosolic double strand DNA (51,52). DNA was known to stimulate immune responses long before it was shown to be genetic material. DNA can be introduced into the cytoplasm by bacterial or viral infection, or due to leakage from nucleus or mitochondria under some pathological conditions. In mammalian cells, cytosolic DNA triggers the production of type I interferons in a sequence-independent way. Mab21D1 or cGAS can respond to the cytosolic DNA and generate a small molecule, cGAMP, to activate the STING protein. Activated STING triggers the downstream signaling cascade including the activation of IRF3 and NF-κB (53).

The mode of how DNA activates the synthesis of cGAS has been recently unraveled by multiple structures of different origins (54-56). Upon dsDNA binding, cGAS is activated through conformational transitions, resulting in formation of a catalytically competent and accessible

nucleotide-binding pocket to generate cGAMP (Figure 1.9) (54). The crystal structure also shows that the generated cGAMP has two types of phosphodiester linkages: G(2',5')pA and A(3',5')pG. Since the single catalytic pocket of cGAS is responsible for the formation of two phosphodiester linkage, it's proposed that the G(2',5')pA is formed first, then the intermediate undergoes the flip-over in order to form the second one. The mixed linkages of this small molecule may be beneficial for its stability since very few RNase target 2', 5'-phosphdiester bond.

The alignment of the Mab21D2 protein sequences from different species shows that this protein may also contain the catalytic motif as Mab21D1 (Figure 1.10). The high sequence identity (~85%) from fish to human may indicate an important physiological function of Mab21D2 in higher organisms.

4. Tables and Figures

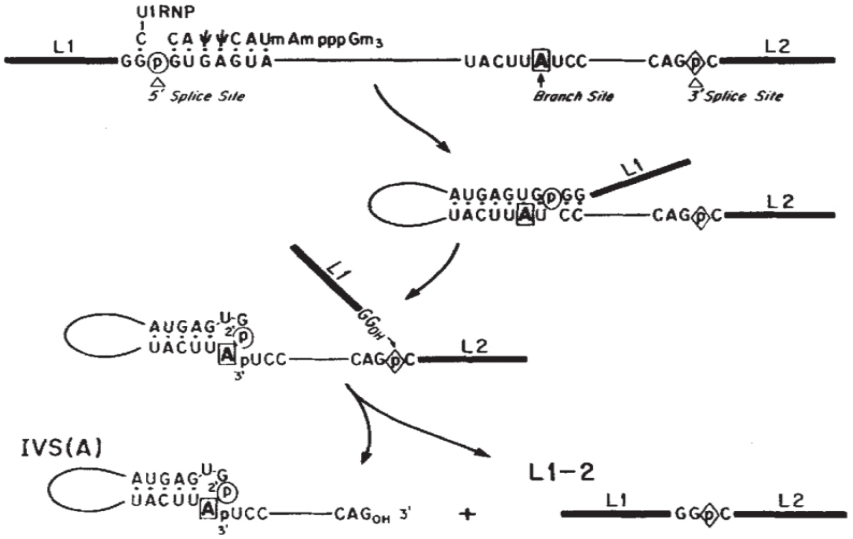

Figure 1.1. Pre-mRNA splicing. Upon the catalysis, the 5'-end intron base pairs with the downstream sequence to form a lariat. The hydroxyl group at branch site attacks the 3'end of exon L1 and causes its release, then 3'-OH of the L1 attacks the 5'-end of exon L2 and ligates with L2 accompanied with the release of lariat intron (4).

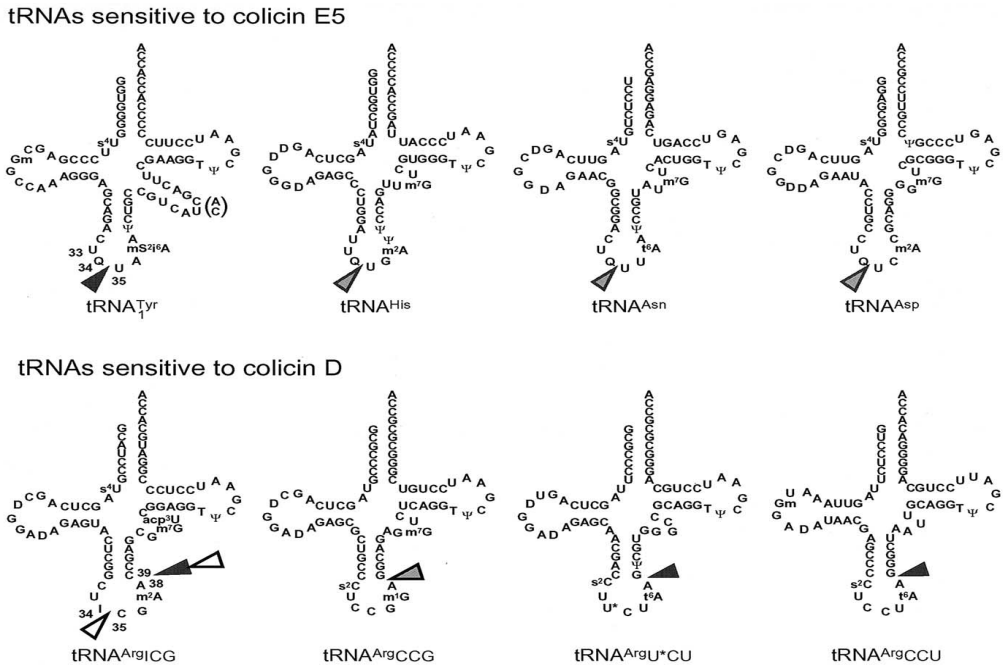

Figure 1.2. tRNA species sensitive to the colicin E5 and colicin D. "Q" stands for queuine. Solid and shade triangles indicate the cleavage sites after in vitro reaction. The open triangles show the cleavage ends observed in colicin D-treated cells (21).

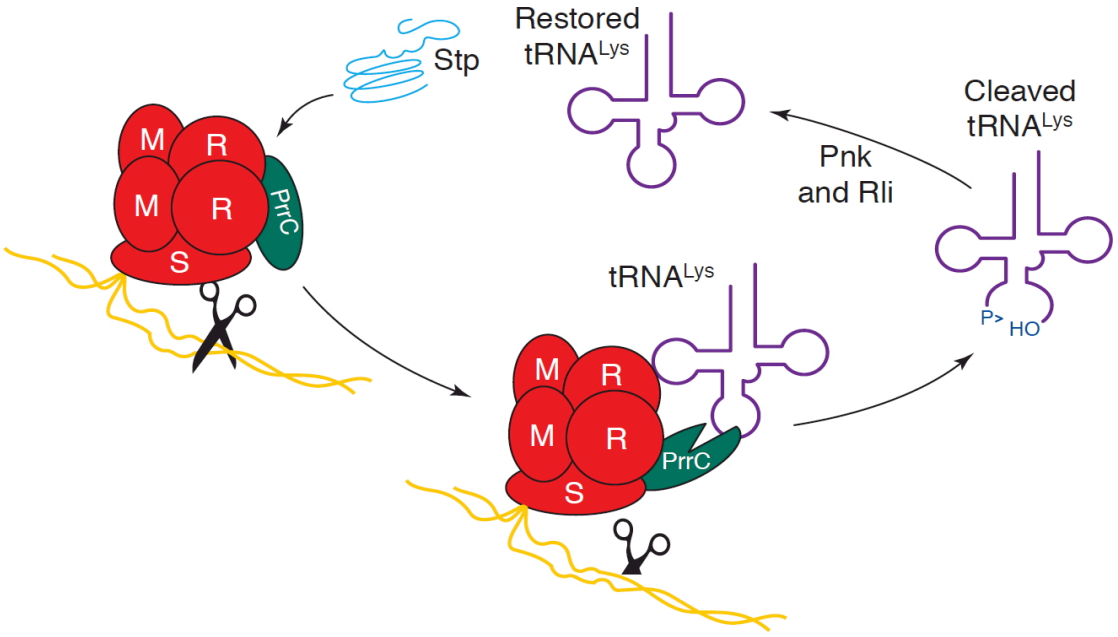

Figure 1.3. *T4*Pnk/Rli escapes from prr restriction. Stp is a participant in host restriction, and it activated PrrC, that cleaves $tRNA^{Lys}$. Pnk and Rli from *T4* can repair the cleaved tRNA through polynucleotide kinase phosphatase and RNA ligase activities (24).

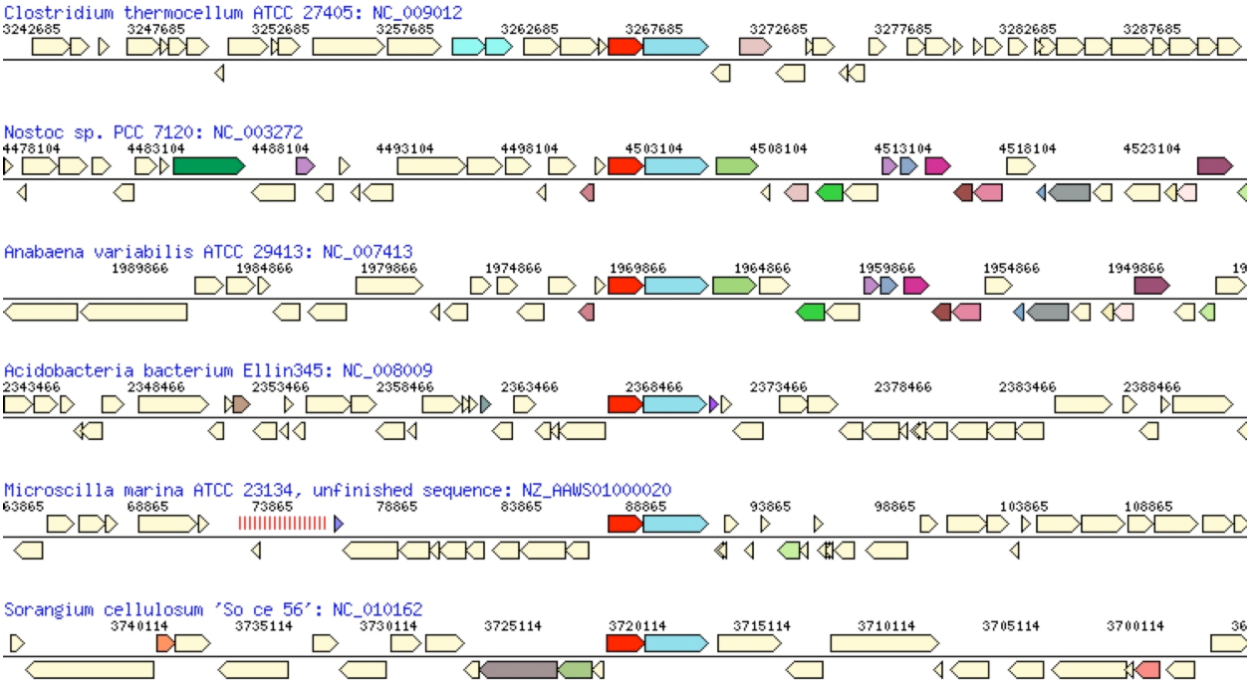

Figure 1.4. Genes surrounding bacterial *hen1* in six representative organisms. Red: the gene encoding a bacterial Hen1. Cyan: the gene encoding a bacterial Pnkp. (Ella Chio mui Chan, thesis dissertation)

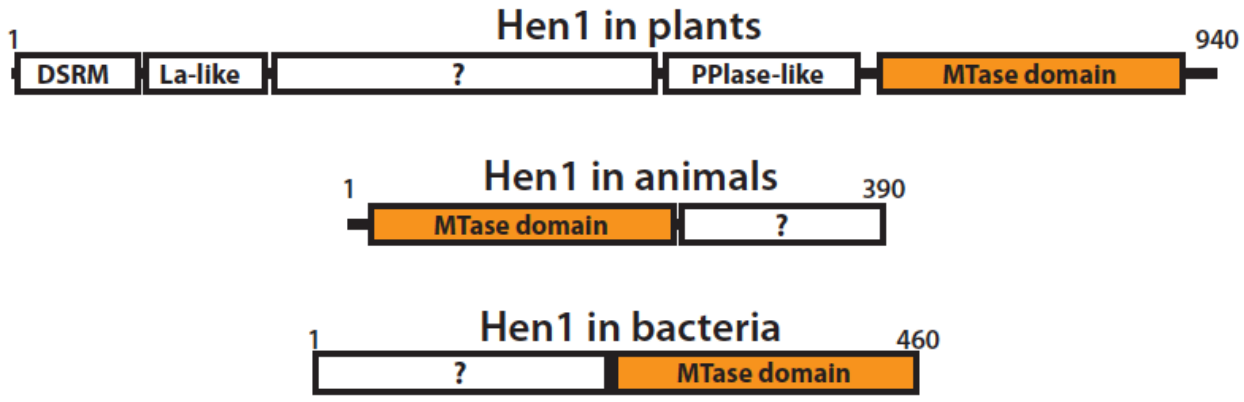

Figure 1.5. Architectures and domain arrangements of Hen1 from plants, animals, and bacteria. DSRM, double-stranded RNA-binding motif; PPIase, peptidylprolyl cis-trans isomerase. The domains labeled with question marks have no significant amino acid sequence similarity to known proteins (35).

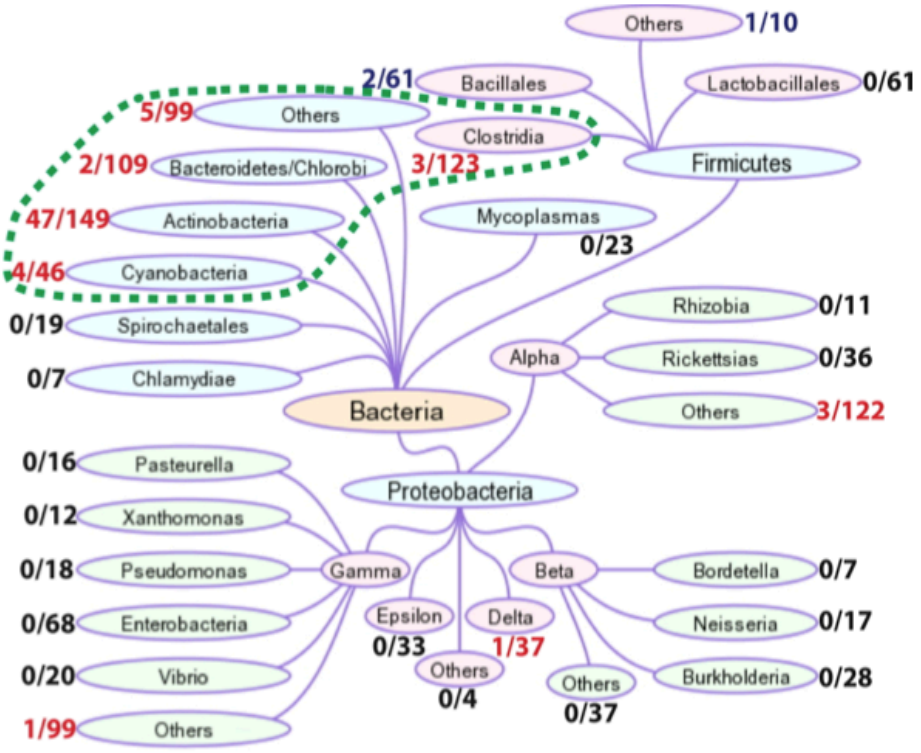

Figure 1.6. Distribution of Pnkp/Hen1 among bacteria. Around 5% of bacterial species with known genomic sequences possess the Pnkp/Hen1 RNA repair system. The distribution is uneven among branches of bacteria, and the majority of bacteria possessing Pnkp/Hen1 belong to Actinobacteria. Within Actinobacteria, all the *streptomyces* species possess Pnkp/Hen1. The two species related to research are *Anabaena variabilis* (*Ava*) belonging to Cyanbacteria branch and *Clostridium thermocellum* (*Cth*) belonging to Clostridia.

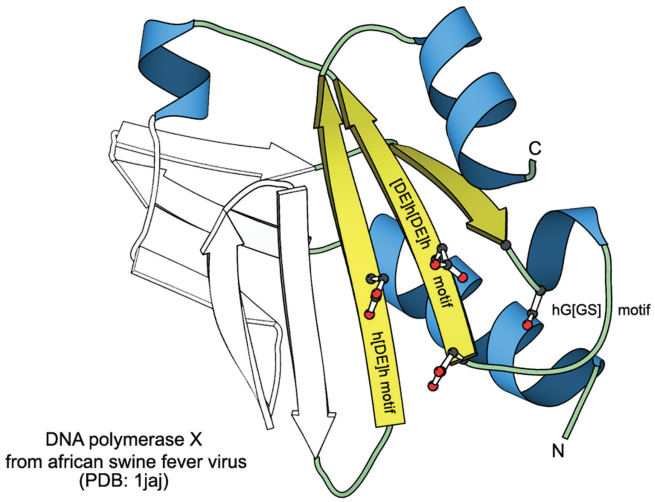

Figure 1.7. NTase fold superfamily structures. The colored part represents the core structure of this superfamily, which is decorated by additional structure (transparent part). Positions of conserved active site motifs involved in catalysis ([DE]h[DE]h, h[DE]h) and substrate binding (hG[GS]) are marked above. Side-chains of critical motifs residues (aspartates and serine) are shown as balls and sticks (42).

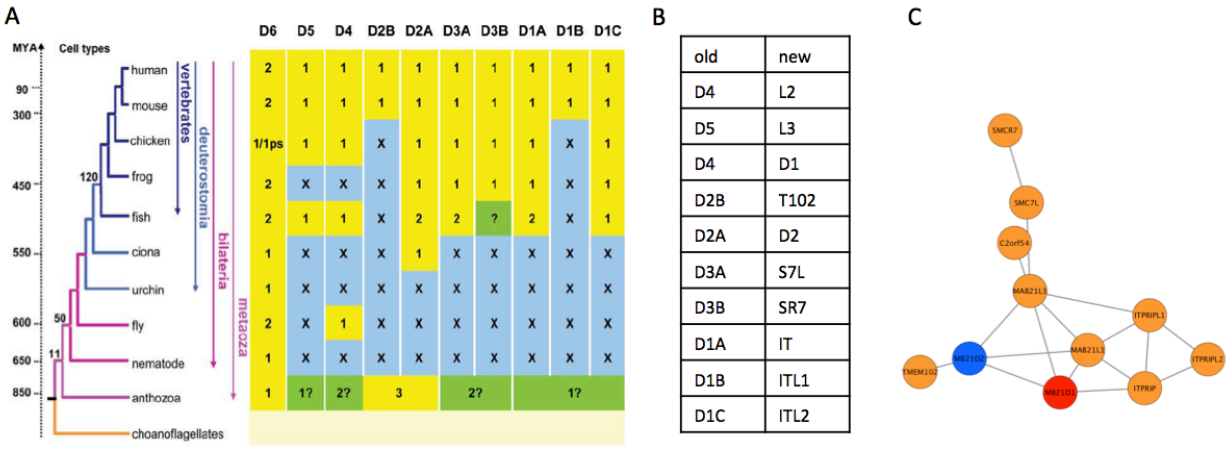

Figure 1.8. Distribution and correlation of Mab21 group members. (A) The left tree is the phylogenetic relationships of representative species. The right table shows whether the corresponding genes exist in those species and the number of copies. "X" indicates the absence of the gene, and question mark indicates the unclassified proteins (48). (B) The comparison of old names and new names used for Mab21 group members; (C) Protein similarity network of Mab21 group members: all the members are collected together when E-cut value is as high as 1e-5. The protein sequences from human are used for making the figure in cytoscape. Among all the members, cGAS (Mab21D1, red node) is recently well studied, and Mab21D2 (blue node) is subject to this thesis research.

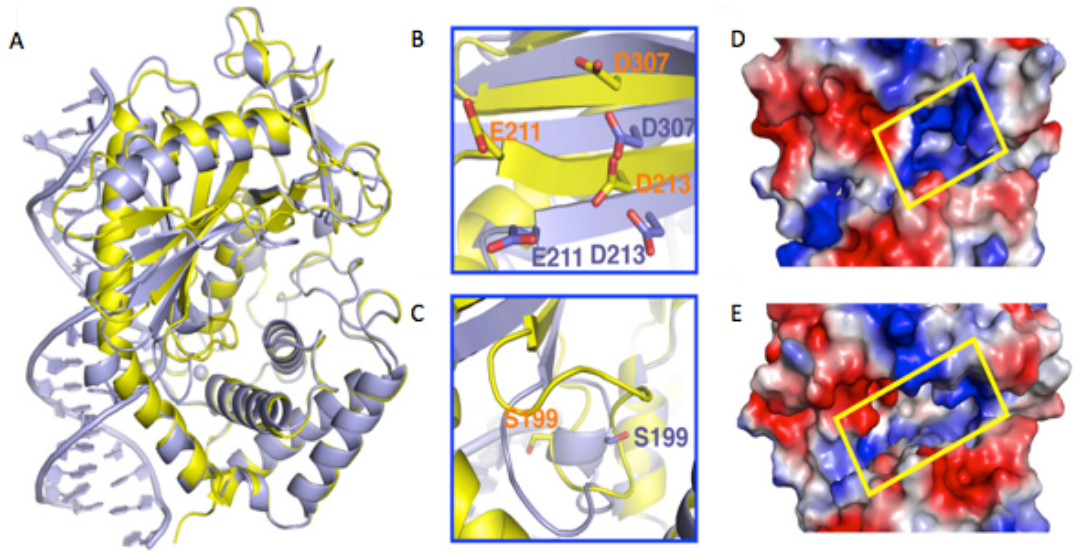

Figure 1.9. dsDNA activates cGAS through conformation change (54). (A) Superposed Crystal structure of cGAS in the free state (yellow) and in the cGAS-dsDNA complex (blue); (B, C) Conformation change of cGAS in the β sheet and catalytic pocket upon the dsDNA binding. (D, E) The protein is in an electrostatic representation (red, acidic; blue, basic; white, neutral); the entrance to catalytic pocket (yellow rectangle) is narrow in the free cGAS (D), while widened in the cGAS with dsDNA bound (E).

```
Oryzias_latipes EDLNRLIQEFSKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDMDLDDLNVYARGTDYDMDFTLLVP
Oreochromis_nil EDLNRLIQEFSKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDMDLDELSIYARGTDYDMDFTLLVP
Trichechus_lati EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Loxodonta_afric EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Odobenus_diverg EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Equus_caballus  EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Nomascus_leucog EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Ailuropoda_mela EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Rattus_norvegic EELNKLIQEFTRHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Mus_musculus    EELNKLIQEFTRHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Otolemur_garnet EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Macaca_fascicul EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Macaca_mulatta  EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Pteropus_alecto EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Cavia_porcellus EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Tursiops_trunca EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Orcinus_orca    EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Saimiri_bolivie EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Callithrix_jacc EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Homo_sapiens-2  EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Homo_sapiens-1  EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Gallus_gallus   EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDIDLDELNVYARGTDYDMDFTLLVP
Xenopus_tropica EELNKLIQEFSKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Ovis_aries      EELNKLIQEFTKHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Monodelphis_dom EELNKLIQEFTRHDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Canis_familiari EMVG-----AQMHDQREYDDQIALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDLDLDELNVYARGTDYDMDFTLLVP
Danio_rerio-1   EELNRLIGEFSKHEQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDMDLDELTVYARGADYDMDFTLLVP
Danio_rerio-2   EELNRLIHEFSKQDQREYDDQRALEIHTAKDFIFSMLGMVQKLDQKLPVANEYLLLSGGVREGVVDMDVDELGVYTRGSDYDMAFTLLVP
Oryzias_lati_28 EQLSAQVSELVLLEQGEFGDQTALEVHTAKDFIFNMLGLVQKVDQRLPVANEYLLLSGGAREGVLDLNPEDLGEYAKGADFDLDFTLLVP
Oreochromis__29 EQLSAQVSELVLLEQGEFGDQTALEVHTAKDFIFNMLGLVQKVDQRLPVANEYLLLSGGAREGVLDLNPEDLGDYAKGADFDLDFTLLVP
Takifugu_rubrip DQLSSQVSELVLLEQRDFGDQTALEVHTAKDFIFNMLGLVQKIDQRLPVANEYLLLSGGAREGVLDLNPEDLGDYAKGADFDLDFTLLVP
Danio_rerio-3   EQLSAQVQELVLLEQGEFGDQTALEVHTAKDFIFNMLGMVQKVDKRLPVANEYLLLSGGAREGVLDLNPEDLGDYAKGVDFDLDFTLLVP
consensus       eelnkliqeftkhdQreydDQrALEiHTAKDFIFsMLGmVQKlDqkLPVANEYLLLSGGvREGVvDldldeLnvYarGtDyDmdFTLLVP
```

Figure 1.10. Conservation of of Mab21D2 protein among vertebrates. The conserved residues are boxed in color, with completely conserved residues in magenta, identical residues in yellow, and similar residues in cyan. The alignment of Mab21D2 protein sequences from representative vertebrates shows around 85% identity, and only the part of alignment containing potential catalytic motif is shown in the figure. The bracket indicates the hypothetic DXD motif that is critical for the catalysis of NTase fold superfamily. This motif is absent in one copy of zebrafish (Danio_rerio) D2 protein.

Table 1.1 Reactions catalyzed by NTase fold family enzymes (42)

REACTION: X + (d)NTP → X-(d)NMP + Y				
Biological process	**Substrate X**	**(d)NTP**	**Product Y**	**Type of substrate**
mRNA polyadenylation	mRNA	ATP	PP_i	RNA
tRNA maturation	tRNA	ATP, CTP	PP_i	
mRNA editing	mRNA	UTP	PP_i	
chromatin remodeling, DNA repair, immunoglobulin gene rearrangement	DNA	dATP, dCTP, dGTP, dTTP	PP_i	DNA
antibiotic resistance	kanamycin, streptomycin	ATP	PP_i	antibiotics
signal transduction (during viral infection)	2'-5'-oligoadenylate	ATP	PP_i	NTP
(p)ppGpp synthesis	GTP	ATP	AMP	
regulation of GS activity (C-terminal NTase domain)	GS	ATP	PP_i	proteins
EXCEPTIONS				
Biological process	**Reaction**			**Type of substrate**
regulation of GS activity (N-terminal NTase domain)	GS-AMP + P_i → GS + ADP			proteins
3'-5'-cAMP synthesis	ATP → 3'-5'-cAMP + PP_i			NTP

Table 1.2 Identified Groups of NTase fold suprfamily (42)

Group	PDB90	COG	KOG	PFAM	Human gene location	ENSEMBL	Diseases	Description
Known superfamily members								
I	2O1P 1Q78 2IKF 2B4V 1R89 1PX5	COG5186 COG5260 COG1746	KOG2245 KOG1906 KOG2277	PF04928	Chr14: 96038471–96103178 Chr2: 60836887–6087960 Chr10: 30641765–30703376 Chr16: 48745280–48820909 Chr5: 78944152–79018227 Chr5: 6767718–6810161 Chr11: 62098756–62115592 Chr9: 88092468–88159198 Chr1: 52661535–52791360 Chr12: 111900657–111933911 Chr12: 111860540–111895433 Chr12: 111829122–111854374 Chr12: 119942478–119961164	ENSP00000216277 ENSP00000238714 ENSP00000263063 ENSP00000350054 ENSP00000369637 ENSP00000230859 ENSP00000278279 ENSP00000365128 ENSP00000360596 ENSP00000342278 ENSP00000228928 ENSP00000202917 ENSP00000257570	Type 1 diabetes	• Poly(A) polymerase • TUTase • DNA polymerase • tRNA nucleotidyltransferase (CCA-adding enzyme) • 2′–5′-oligoadenylate synthetase • TRF4/5 nucleotidyltransferase
II	1MIV 1VFG 1OU5	COG0617	KOG2159	PF01743	Chr3: 3107800–3131722	ENSP00000354999		• tRNA nucleotidyltransferase • Poly(A) polymerase
III	2PBE			PF04439				• Streptomycin adenylyltransferase
IV	1KAN							• Kanamycin nucleotidyltransferase
V	1WOT 1NO5 2RFF 1YLQ	COG1669 COG1708		PF01909				• Predicted nucleotidyltransferase
VI	1BPD 1XSL 1JMS 2IHM 1JAJ	COG1796	KOG2534		Chr10: 103328629–103338004 Chr10: 98054075–98088290 Chr7: 44078374–44088607 Chr8: 42315131–42348482	ENSP00000359187 ENSP00000360216 ENSP00000242248 ENSP00000265421		• DNA polymerase
VII	2GA9			PF03296				• Poxvirus poly(A) polymerase
VIII	1V4A	COG2844 COG1391 COG2905		PF03710 PF03445				• Glutamine synthetase adenylyltransferase • Putative nucleotidyltransferase
IX	1VJ7 2BE3	COG0317 COG2357	KOG1157	PF04607				• Guanosine polyphosphate pyrophosphohydrolase/synthetase
X	2EWR							• Predicted nucleotidyltranferase
XI	2NRK	COG2320		PF04229				• Predicted nucleotidyltransferase
XII		COG2413						• Predicted nucleotidyltransferase
XIII		COG4914						• Predicted nucleotidyltransferase
XIV		COG3541						• Predicted nucleotidyltransferase
XV		COG3072		PF01295				• Adenylate cyclase
XVI		COG1665						• Predicted nucleotidyltransferase

Table 1.2 Identified Groups of NTase fold suprfamily (42) (cont.)

XVII	COG4849						• Uncharacterized conserved protein
XVIII	COG5397						• Uncharacterized conserved protein
XIX	COG2253		PF08843				• Uncharacterized conserved protein
XX	COG3575		PF06042				• Uncharacterized conserved protein
XXI		KOG2054	PF03813	Chr9: 33451354–33463941	ENSP00000368784		• NRAP
XXII		KOG3792	PF07528	Chr19: 3755022–3785924	ENSP00000262961		• NFAT
		KOG3793		Chr5: 32390214–32480601	ENSP00000371560		• Predicted nucleotidyltransferase
				Chr9: 124926811–125070676	ENSP00000362742		
				Chr19: 10642282–10656141	ENSP00000250241		
				Chr1: 151900905–151910148	ENSP00000355011		
XXIII	COG3475		PF06828	Chr9: 107360232–107443220	ENSP00000223528	Fukuyama-type congenital muscular dystrophy	• LICD
			PF04991				• Fukutin-related proteins
				Chr19: 51943121–51953581	ENSP00000326570	Congenital muscular dystrophy type 1C	
XXIV		KOG3963	PF03281	Chr2: 241474139–241484087	ENSP00000373586		• Mab-21
				Chr4: 151722699–151725262	ENSP00000324701	Neurologic disorders	
				Chr13: 34946319–34948830	ENSP00000369251	Neurologic disorders	
				Chr3: 193997301–194118644	ENSP00000314252		
				Chr10: 106061884–106088152	ENSP00000350915		
				Chr16: 19033206–19034943	ENSP00000370849		
				Chr22: 38228230–38244079	ENSP00000327124		
				Chr2: 96355688–96357806	ENSP00000355121		
				Chr17: 18104594–18109818	ENSP00000323591		
				Chr6: 74179959–74218720	ENSP00000296913		
				Chr17: 7279486–7281722	ENSP00000315387		
				Chr1: 116455899–116479384	ENSP00000358512		
XXV		KOG3852	PF07984	Chr1: 117950079–117972517	ENSP00000358458		• FAM46
				Chr6: 82512166–82519210	ENSP00000358771	Retinal pigmentosa	
				Chr1: 27204098–27212049	ENSP00000351491		
				ChrX: 79477659–79587466	ENSP00000342730		
XXVI		KOG2986	PF09139	Chr3: 11806920–11863355	ENSP00000273037		• MMP37

CHAPTER 2: BIOCHEMICAL AND STRUCTURAL STUDIES OF ACTIVE RNA LIGASE PNKP-C/HEN1-N INVOVLED IN BACTERIAL RNA REPAIR

1. Introduction

RNAs in living organism experience constant damages, and some of the damages are spontaneous, which belong to normal RNA metabolism (3-7). However, RNA damage could also be specific. To respond to the cellular stress or distinguish self from nonself, some microbes can produce specific endoribonuclease toxin to target essential RNAs involved in protein translation. For example, colicin D and E5 are a class of antibiotics produced by a certain strain of *E coli*, and are secreted to target specific tRNA and cause the cell death (20-22). Most ribotoxins employ a transesterification mechanism (57) to cleave essential RNA and generate broken ends with 5'-OH and 2', 3'-cyclic phosphate.

To counter the purposeful RNA damages carried out by ribotoxins, there are RNA repair systems available in some organisms. The first RNA repair system was found in *T4* phage, and two proteins Pnkp (polynucleotide kinase-phosphatase) and Rli (RNA ligase) were involved (27). *T4*Pnkp is a bifunctional enzyme that removes the 2', 3'-cyclic phosphate and phosphorylates 5'-OH, while *T4*Rli can ligate the two processed ends and restore the original form of RNA (Figure 2.1). Our lab discovered that a bacterial RNA repair system consisting of two proteins Pnkp and Hen1. The bacterial Pnkp has previously been shown to have kinase, phosphatase and adenylyltransferase activities, which are the hallmarks for RNA repair (30). But it alone can only process the broken RNA but not seal it. Within the same operon, there is always another gene encoding bacterial Hen1 (Hua enhancer) upstream of Pnkp (Figure 1.4). Hen1 in eukaryote is SAM-dependent 2'-O-methyltransferase to modify RNA (36-40). Bacterial Pnkp and Hen1 are shown to form heterotetramer in vitro and carry out RNA repair. Moreover,

Pnkp/Hen1 can also add a methyl group at the ligation junction of RNA and prevent the repeated damages (Figure 2.2) (34). To ensure the opportunity of 2'-O- methylation by bacterial Hen during RNA repair and maintain the quality of repaired RNA, Pnkp/Hen1 has evolved to require the participation of Hen1 in RNA ligation, even Pnkp C-terminal adenylyltransferase possess all the signature motifs of RNA ligase (Figure 2.3) (31).

We found that C-terminal adenylyltransferase of Pnkp can interact with N-terminal domain of Hen1 to form heterodimmer, which has the RNA ligation activity (Figure 2.5). We crystallized *Cth*(*Clostridium thermocellum*)Pnkp-C/Hen1-N and *Cth*Hen1-N homodimmer, while another group solved the structure of *Cth*Pnkp-C (31). We proposed molecular basis of ligase activation: formation of heterodimer creates a ligation pocket for the RNA substrates (58).

In addition to the structure, we also carried out functional studies of this complex, including the protein adenylylation reaction in crystal, RNA adenylylation in solution and mutagenesis study (58). All the results together provide insight into the mechanism of how Hen1 activates RNA ligase activity of Pnkp during RNA repair.

2. Methods and Materials

2.1. Cloning, overexpression and purification of recombinant protein

DNA fragments corresponding to the C-terminal half of *Cth*Pnkp (residues 444-870) and N-terminal half of *Cth*Hen1 (residues 1-230) were PCR amplified from the genomic DNA of *Clostridium thermocellum* and inserted into the pETDuet-1 vector. The truncation *Cth*Pnkp-C-ΔL (residues 479-870) and *Cth*Hen1-N- ΔL (residues 1-199) were constructed in the same way. The constructed vectors were transformed into BL-21(DE3) for protein expression. 5ml overnight culture was inoculated 1 liter Lysogeny broth (LB) containing 0.05 mg/ml ampicillin.

Cells were cultured at 37°C until A600nm reached 0.5 and then cooled down to 18°C, induced overnight with 0.5mM IPTG. Cells were harvested by centrifugation at 4000g for 30 minutes, and the pellet was stored at -80°C.

The cell pellet was thawed and resuspended in the lysis buffer [20 mM Tris-HCl (pH 8.0), 10 mM NaCl, 1 mM PMSF, 1mM DTT] and lysed using a French press. The cell lysate was centrifugated at 14000rpm and the proteins were purified to homogeneity by using FPLC. The purifications of *Cth*Pnkp-C and *Cth*Pnkp-C-ΔL were as follows: the supernatant passed through a DEAE anion exchanger column equilibrated by DEAE A buffer (20 mM pH 8.0 Tris-HCl, 10 mM NaCl, 2% glycerol, 1mM DTT). The proteins were eluted by gradually increasing the percentage of DEAE B buffer (same as DEAE A buffer, but containing 1M NaCl) to 100%. Fractions containing the protein were pooled and diluted with DEAE A buffer and applied to Heperin column equilibrated with DEAE A buffer. The bound proteins were eluted with the DEAE B buffer, diluted by DEAE A buffer and applied to Mono Q anion exchange column. Proteins were eluted from Mono Q column by gradually increasing the percentage of DEAE B buffer and loaded into Superdex 200 size exclusion column equilibrated with a Gel filtration buffer (10 mM HEPES (pH 7.0), 200 mM NaCl, 1 mM DTT, 2% Glycerol). Proteins were eluted by Gel filtration buffer and concentrated by 10K cut centricon (Millipore). The protein concentration was measured through Biorad reagent.

The purification of *Cth*Hen1-N was similar to *Cth*Pnkp-C, except that Heparin column was skipped. *Cth*Hen1-N-ΔL was shown to be insoluble and no protein was got from the same method.

*Cth*Hen1-N and *Cth*Pnkp-C were mixed with 1:1 molar ratio for 1 hour at 4°C, and then applied to Superdex 200 size exclusion column. The fraction containing heterodimmer was

collected and concentrated for further study.

2.2. Crystallization, data collection and structure determination

Crystals of the *Cth*Pnkp-C/Hen1-N heterodimer were grown by the hanging-drop vapor diffusion method at 30 °C. The purified *Cth*Pnkp-C/Hen1-N heterodimer (~6 mg/mL) was mixed with a reservoir solution containing 16% PEG4000, 0.5 M NaCl, 25 mM CaCl2, 5% methanol, 50 mM Hepes (pH 7.5), and 50 mM Tris·HCl (pH 8.0). Thin plate crystals appeared after 2–3 d. Crystals of the *Cth*Hen1-N homodimer were grown in a similar manner, except that the growing temperature was 18 °C and the reservoir solution contained 15% PEG400, 0.1 M NaCl, and 0.1 M MES (pH 6.5). Crystals were soaked briefly in a cryoprotecting solution containing all of the components of the reservoir solution supplemented with 25% (vol/vol) glycerol, and the cryoprotected crystals were then mounted in a nylon loop and flash-frozen in liquid nitrogen. To obtain the crystal structure of the *Cth*Pnkp-C/Hen1-N/AMP ternary complex, a crystal of *Cth*Pnkp-C/Hen1-N heterodimer was soaked with 5 mM ATP and 5 mM Mg^{2+} in the cryoprotecting solution overnight before crystal mounting and data collection. Data were collected at 21-ID beam lines at the Advanced Photon Source (APS) and processed using the HKL2000 program.

Phase for the structure of *Cth*Hen1-N homodimer was determined based on a SeMet SAD data using the Phenix program. A partial model was automatically built using Phenix (68). The remaining model was manually built using the Coot program (69). Refinement was carried out using Phenix. The structure of one monomer of *Cth*Hen1-N homodimer, together with the structure of ligase domain of *Cth*Pnkp (PDB: 3TY8, the insertion module was deleted) (31), was used as the initial search model for solving the structure of *Cth*Pnkp-C/Hen1-N heterodimer.

Two copies of each were found using the molecular replacement method in Phenix. The model building and the structural refinement of the *Cth*Pnkp-C/Hen1-N heterodimer and the *Cth*Pnkp-C/Hen1- N/AMP ternary complex were analogous to the procedures for the structure of the *Cth*Hen1-N homodimer.

2.3. RNA preparation

2.3.1 DNA template purification

To generate E. coli $tRNA^{Arg}$ – ΔT, two strands of systhetic DNA purchased from DNA IDT served as template for in vitro transcription. Both DNA strands were purified by 8% denaturing polyacrylamide gel electrophoresis (DPAGE): DNA were dissolved in TE buffer (pH8.0), mixed with equal volumn of DPAGE loading buffer (10 mM of EDTA, pH 8.0, in 100% formamide) before loaded into gel. The desired DNA bands were located under Ultraviolet light, and extracted with elution buffer (0.5 M $NH_4C_2H_3O_2$, 1 mM EDTA). The DNAs were precipitated under -80°C for 30 minutes after being mixed with 1/10 volume of 3M $NH_4C_2H_3O_2$ and 3 volumes of icy cold ethanol. DNA was recovered by centrifuged at 13,200 rpm for 30 minutes. DNA pellet was resuspended into TE buffer and adjusted the concentration to 100uM.

2.3.2 In vitro transcription

The two strands of DNA were mixed with 1:1 molar ratio and annealed by heating to 95 °C for 3 min and then slowly cooling to 25 °C. The in vitro transcription reaction was carried out in 4 ml volume containing 5 mM NTP mixture (ATP, CTP, GTP, and UTP), 40 mM Tris-HCl (pH 7.9), 6 mM MgCl2, 50 mM dithiothreitol (DTT), 2 mM spermidine, 1 μM DNA template and 1mg purified T7 polymerase. After incubation overnight at 37°C, the RNA molecules were

gel purified the same way as DNA purification. The same protocol was employed for the preparation of ^{33}P-internal radiolabeled tRNA^{Arg} – ΔT with the exception that the scale is 40ul and 50 μCi alpha-^{33}P radiolabeled GTP was added in the reaction.

2.3.3 RNA cleavage and modification

tRNA^{Arg} – ΔT was ajusted to 100uM final concentration in the buffer containing 25 mM HEPES (pH 7.0), 25 mM NaCl, 5 mM $MgCl_2$, 0.05 mg/ml BSA, heated at 95°C for 5 minutes and then cooled down on ice for the formation of secondary structure. tRNA cleavage was carried out by being incubated with 25 uM ColicinD at room temperature for 2 hours, and the products (38mer and 22mer) were identified and purified by 15% gel. The cleaved products were labeled as 5'-half-P (there is 2', 3'- cyclic phosphate at the 3'-end of 5'-half RNA), 3'-half. The radiolabeled RNA was cleaved in the same way, and the cleaved products were labeled as 5'-half-P* and 3'-half*.

To prepare RNA for ligation assay, both 5'-half-P* and 3'-half* need to be treated with *Cth*Pnkp-N (only contains kinase and phosphatase domain) to get 2', 3'-cyclic phosphate removed and 5'-OH phosphoryalted. 100uM cleaved internal radiolabelled tRNA ^{Arg} – ΔT was incubated with 10uM *Cth*Pnkp-N in the reaction buffer (25 mM Tris·HCl, pH 8.0, 50 mM KCl, 2.5 mM $MgCl_2$, 0.05 mM EDTA, and 5 mM DTT, 0.5 mM ATP and 0.25 mM Mn^{2+}) at 45°C for 45 minutes. The modified RNA was purified by phenol extraction followed by ethanol precipitation. The modified product was labeled as 5'-half*, 3'-half-P*.

To prepare RNA for the step-2 ligation assay, 3'-half RNA (22mer) need to be phosphorylated at 5'-OH. 100 uM of cold 22mer was incubated with 10 uM *Cth*Pnkp-N in the reaction buffer (25 mM Tris·HCl, pH 8.0, 50 mM KCl, 2.5 mM $MgCl_2$, 0.05 mM EDTA, and 5

mM DTT, 0.5 mM ATP and 0.25 mM Mn^{2+}) at 45°C for 45 minutes. The modified RNA was purified by phenol extraction followed by ethanol precipitation.

2.4. Biochemical assay

For all enzymatic assays, the required proteins were mixed and preincubated at 25 °C for 1 h before use to promote formation of the complex (as needed).

2.4.1 RNA ligation assay

5'-half*/3-half-P* were mixed with equal molar ratio in the repair buffer (containing 25 mM Tris·HCl, pH 8.0, 50 mM KCl, 2.5 mM $MgCl_2$, 0.05 mM EDTA), annealed by heating to 95°C for 3minutest and slowly cooling down to room temperature. 1uM annealed RNA was incubated with 0.5uM enzyme in the same reaction buffer plus 5 mM DTT, 0.5 mM ATP and 0.25 mM Mn^{2+}. Reaction was stopped after 45 minutes incubation by phenol extraction, and followed by ethanol precipitation. The RNA pellet was dissolved in 20 μl of DPAGE loading buffer and analyzed by 15% DPAGE. The radioactivity of RNA was quantified by using a PhosphorImager system (Molecular dynamics).

2.4.2 Step-2 ligation assay

Enzyme (1 μM) was incubated with 1 μCi of ^{32}P-alaph-ATP in the repair buffer at 45 °C for 10 min. ATP (0.1 mM) was then added, and the sample was incubated at 45 °C for another 10 min. 2uM RNA substrate (3'-half-P, or annealed 5'-half-P/3'-half-P) was added into the reaction buffer and mixed with the pretreated enzyme, and the mixture was incubated at 45 °C for 10 minutes. The reaction was stopped with phenol extraction followed by ethanol precipitation, and the recovered RNA was analyzed analogous to the protocols described in the

RNA ligation assays.

2.4.3 Step 2+3 ligation assay

The enzyme–AMP covalent intermediate was formed the same as in the assay for step 2, except incubation with ^{32}P-alpha-ATP was omitted. To phosphorylate the 5'-end and dephosphorylate the 3'-end of the RNA substrate, the annealed 5'-half-P*/3'-half* (2 μM) was incubated with *Cth*Pnkp-N (2 μM) in the reaction buffer at 45 °C for 15 min in the presence of 0.1 mM ATP and 0.25 mM Mn^{2+}. The pretreated enzyme and RNA substrate were then combined in equal volume, and the combined sample was incubated at 45 °C for 20 min to carry out steps 2 + 3. The reaction was stopped with phenol extraction followed by ethanol precipitation, and the recovered RNA was analyzed in a manner similar to the protocols described in the RNA ligation assays.

2.5. Functional analysis of mutants

The expression vectors for single-alanine mutants were created based on the expression vectors of *Cth*Hen1-N and *Cth*Pnkp-C –ΔL using QuickChange Method. There were 5 single mutants for *Cth*Hen1-N: L61A, R63A, L74A, Y77A H194A and 11 single mutants for *Cth*Pnkp-C –ΔL: W611A, K614A, L618A, Q622A, Y623A, Y689A, Y798A, K824A, E862A, D867A, R869A. The construct of deletion mutants was similar to the original construct, except the first residue of Hen1-N and the last residue of Pnkp-C- ΔL were deleted respectively. All the mutated proteins were overexpressed and purified in the same way as the native proteins.

3. Results and Discussions

3.1. Reconstitution of an active RNA ligase *Cth*Pnkp-C/Hen1-N heterodimmer *in vitro*

3.1.1 Formation of *Cth*Pnkp-C/Hen1-N heterodimmer *in vitro*

The recombinant *Cth*Pnkp-C (residues 444-870; 49kDa) and *Cth*Hen1-N (residues 1-230; 26kDa) were overexpressed in *Escherichia coli* and purified to homogeneity. Size-exclusion chromatography of the purified proteins showed that *Cth*Pnkp-C was monomer in solution while *Cth*Hen1-N was homodimer. (Figure 2.4A, green line and red line). The incubation of equal moles of *Cth*Pnkp-C and *Cth*Hen1-N resulted in the formation of heterodimer instead of heterotetramer (Figure 2.4A, black line). We proposed that the interfaces for *Cth*Hen1-N homodimer and *Cth*Pnkp-C/Hen1-N heterodimer were overlapped. The structure supports our hypothesis, and the detailed evidence would be discussed later.

3.1.2 Reconstitution of RNA ligation activity

To carry out functional analysis, we chose $tRNA^{Arg}$ – ΔT cleaved by ribotoxin colicin D since the preliminary data showed that cleaved $tRNA^{Arg}$ – ΔT was the best substrate of all we have tested (Figure 2.4B). The cleaved RNA substrate was pretreated with *Cth*Pnkp-N, so the 2', 3'- cyclic phosphate was already removed and 5'-OH was phosphorylated. The two RNA fragments were ready for ligation by ligase. The data (Figure 2.4C) showed that, the full-length *Cth*Pnkp and *Cth*Hen1 were able to ligate the RNA. The combination of *Cth*Pnkp-C and *Cth*Hen1-N, but neither one alone, could carry out RNA ligation. We also tested a shorter version *Cth*Pnkp-C-ΔL (residues 479–870; 45 kDa, it lacked the linker region compared to *Cth*Pnkp-C) for RNA ligation, and it shows higher ligation activity.

3.2. Overall structure

We crystallized the *Cth*Pnkp-C/Hen1-N heterodimer and solved the structure at 2.6 Å, and also obtained ternary structure of heterodimer covalent bound to an AMP. The ternary structure also has high resolution and does not change conformation much upon AMP binding, so it's used for detailed structural analysis. In addition, we also solved the *Cth*Hen1-N homodimer structure, while another group solved the structure of *Cth*Pnkp-C (31).

3.2.1 Overall structure of *Cth*Hen1-N homodimer

The fold of *Cth*Hen1-N can be depicted as a seven-stranded antiparallel β-sheet stacked by two layers of helices (Figure 2.5). We tentatively assigned the β-sheet together with the first layer of helices as the ligase-activaing domain (color in red and pink in Figure 2.5) and the second layer of helices (color in yellow and orange in Figure 2.5) as part of the linker region connecting to the C-terminal missing methyltransferase. It has been showed that the linker region is related to the solubility of Hen1-N, so we were not able to delete the linker region to analyze its activity. The dimerization of Hen1-N is based on the interaction of the β-sheet from two monomers, which is mainly hydrophobic. The structure of Hen1-N homodimer also has some unstructured parts, especially the loops connecting the β strands.

3.2.2 Overall structure of *Cth*Pnkp-C/Hen1-N/AMP

The ligase domain of *Cth*Pnkp, as previously reported by Shuman and coworkers, is composed of an NTase module, an insertion module and a C-terminal module (31). In our structure, Pnkp-C also contained extra 35 amino acids at its N-terminal (Figure 2.6A). We assigned the extra part as linker region connecting the N-terminal missing kinase and

phosphatase domains (30). *Cth*Hen1-N and *Cth*Pnkp-C make extensive contacts with each other, resulting in a total 3,850 $Å^2$ solvent accessible surface area buried in the dimmer interface (Figure 2.6E). Generally, the ligase-activing domain of Hen1-N, and insertion domain, the C-terminal additional domain of Pnkp-C are responsible for the heterodimer formation (Figure 2.6B, C). Since the β-sheet of Hen1-N is the common interface for *Cth*Hen1-N homodimer and *Cth*Pnkp-C/Hen1-N heterodimer (compare Figure 2.6C with Figure 2.4), so it is understandable that heterodimer instead of heterotetramer form if *Cth*Pnkp-C and *Cth*Hen1-N are incubated together.

3.3. Conformation change upon the heterodimer formation

Significant structural changes take place in both *Cth*Hen1-N and *Cth*Pnkp-C upon heterodimer formation.

3.3.1 Conformation change of *Cth*Hen1-N

Compared to *Cth*Hen1-N homodimer, *Cth*Hen1-N from *Cth*Pnkp-C/Hen1-N heterodimer maintains main body of the structure (β-sheet with two layers of α-helix) (Figure 2.7A). However, the loops connecting the neighboring β-strand could only be seen in the heterodimer. To facilitate the analysis, we depict the Hen1-N structure as a hand. The β-sheet is like palm while the loops connecting the neighboring β-strand are like thumb and fingers. The palm of *Cth*Hen1-N is the major interface for the interaction to *Cth*Pnkp-C.

3.3.2. Conformation change of *Cth*Pnkp-C

In the absence of *Cth*Hen1-N, the insertion domain of *Cth*Pnkp-C is conformationally flexible, as demonstrated by different folds and orientations in the structure of *Cth*Pnkp-C/ATP

and *Cth*Pnkp-C/AMP complex solved by Shuman's group (31). Insertion domain in the *Cth*Hen1-N/Pnkp-C is fixed in the similar orientation as that in *Cth*Pnkp-C/AMP, but the relative position of the two helices that constitute the insertion module are opposite to each other. Moreover, the side chains of six conserved residues located at the bottom of the two helices are scattered in *Cth*Pnkp-C/AMP structure, but orientated in one direction in *Cth*Hen1-N/Pnkp-C structure (Figure 2.7B). The interaction between *Cth*Hen1-N and *Cth*Pnkp-C also occurs at the C-additional module of *Cth*Pnkp-C, resulting in a small but significant change. Particularly, the fragment corresponding to residues 798 to 815 shows obvious backbone shift, and leads to the orientation change of some conserved residues, such as D806, R815 (Figure 2.7C). The conserved residues in insertion module and C-additional module of *Cth*Pnkp-C are subject to the later mutation study.

3.4. Detailed interaction between *Cth*Pnkp-C and *Cth*Hen1-N

The palm (β-sheet) of the ligase-activating domain only interacts with the insertion module of *Cth*Pnkp-C. This interaction is mainly hydrophobic, and a salt bridge between conserved residues D55 and R671 (Figure 2.8, 2.9) is also important for the recognition (Figure 2.10A).

The interaction between the thumb (the loop connecting two β-strands) and insertion module are mainly salt bridge and hydrogen bond contributed by D153 and E154 from *Cth*Hen1-N, R631, Y686 and R687 from *Cth*Pnkp-C. Moreover, W159 from thumb stack on the opposite side of the insertion module where the important conserved residues are located (Figure 2.10B).

Unlike the palm and thumb, finger 2 and finger 3 (the loop connecting other β-strands) of the ligase-activating domain make contacts with both insertion module and C-addition module

(Figure 2.10C). In the center of this three-way interaction, there is extensive hydrogen bonding network contributed by a water molecule as well as the residues from fingers (L54, P82 and Y83), insertion module (R675) and C-addition module (D806). Moreover, a salt bridge between the side chains of D57 and R815, and a π-charge stacking between the side chain of Y77 and R869, further stabilize the interaction between the ligase-activating domain and C-addition module.

Remarkably, the C-terminal motif (DPRL) of Pnkp-C is highly conserved among different species (Figure 2.9). In the *Cth*Hen1-N/Pnkp-C structure, we found that the main chain of R869 and L870 form the salt bridges with R794 and K792, which are members of motif V of NTase module. The role of C-terminal motif in the stabilization of the functional residues may help to explain its strict conservation.

3.5. Ligation pocket formation

Formation of the *Cth*Pnkp-C/Hen1-N heterodimer creates a deep pocket with the covalent AMP at the bottom. The pocket is divided by the side chain of R565, which belongs to motif Ia of the NTase module. The NTase module essentially provides the entire floor of the ligation pocket, but it only contributes about 25% of the wall. The C-additional module provides ~50% of the wall, and the remaining ~25% of the wall is from the joint contribution of the insertion module and the ligase-activating domain (Figure 2.11, panel A). The pocket is predominantly positively charged, which is consistent with its proposed role in accommodating negatively charged RNA substrates.

Guided by the position of the covalently bound AMP and the structure of human DNA ligase I in complex with DNA (63, 64), we docked two strands of 4nt RNA into the ligation

pocket. In the docking model, the two strands of RNA fit nicely into the pocket (Figure 2.11, panel C, D), with most of negative phosphate group facing to the positively charged wall. In the pocket, the covalently bound AMP is readily to be transferred to the 5'-phosphate of 3'-half RNA, while 3'-OH of 5'-half RNA is nearby for the sealing process (Figure 2.12). The pocket is around two nucleotides deep according to the model, and moreover, no sequence specificity is implied. The docking model may provide molecular basis for that this repair complex may have broad substrate spectrum (41).

3.6. RNA ligation steps assay

A classical DNA/RNA ligation reaction contains three enzymatic steps (59-62): 1) ligase reacts with ATP to form a ligase-AMP covalent intermediate; 2) AMP is transferred to the 5'-phosphate in 5'-half RNA; 3) ligase catalyzes nucleophilic attack of the 3'-OH in 3'-half RNA to the adenylylated 5'-phosphate in 5'-half RNA in order to seal the substrate, which is accompanied by the release of AMP (Figure 2.13A) (59).

We carried out the step-1 ligation in the crystal of *Cth*Pnkp-C/Hen1-N by soaking the crystal with ATP and Mg^{2+} before data collection. In the ternary structure of *Cth*Pnkp-C/Hen1-N/AMP, K531 covalently links to an AMP through the phosphate group. There is also an Mg^{2+} in the active site, which is coordinated by the phosphate from AMP and 5 water molecules nearby (Figure 2.12).

Previous study shows that Pnkp alone is able to carry out the step-1 ligation (31), but Hen1 is required for completing ligation function (34). We carried out the following experiment to explore whether Hen1 started to facilitate ligation from step 2 (Figure 2.13B). We incubated radioactive P^{32}-alpha-ATP with enzyme and RNA substrate, checked whether RNA-AMP could

be formed. We found that *Cth*Pnkp-C-ΔL could only adenylylate a small amount of 3'-half-RNA (with 5'-phosphate) in the absence of Hen1-N. However, Hen1-N could enhance the activity by about 2000 fold (Figure 2.13C). Moreover, the presence of 5'-half-RNA (its 2', 3'-cyclc phosphate would prevent the completion of ligation) reduced the efficiency of the step-2 ligation for some unknown reason.

3.7. Mutants study

To provide detailed view of the roles of conserved residues in RNA ligation, we carried out mutation and deletion study (Figure 2.14). We focus on the residues beyond the NTase module of *Cth*Pnkp and their effects on step-2 and step-3 ligation, because mutagenesis of the conserved residues in the NTase module of *Cth*Pnkp and their effects on step-1 ligation have already been reported by Shuman and coworkers (31). According to the sequence alignment (Figure 2.8, 2.9), we chose six sites (M1, L61 R63, L74, Y77, H194) in the ligase-activating domain (Figure 2.14A), six (W611, K614, L618, Q622, Y623 and Y689) in insertion module (Figure 2.14B) and six amino acids (Y798, K824, E862, D867, R869A and L870) in C-additional module (Figure 2.14C). We have deleted the first amino acid (M1) in Hen1 and last amino acid (L870) in Pnkp, and carried out alanine mutation for the rest of 16 chosen amino acids. Except for M1, all the other 17 sites chosen are part of or near the ligation pocket. We have tested the substrate adenylylation and phosphodiester bond formation for all the mutations, and compared the effect with wide type protein.

Overall, the effect of most mutations on catalysis is modest, presumably because the sites chosen are not directly related to classical motifs for ligase. We classify these 18 mutations or deletions into three groups according to their effects on the enzymatic activity.

The first group contains eight mutants that have significant reduced enzymatic effect, which is defined as less than one third of control (Figure 2.14D). Within the first group, alanine mutations of some aromatic residues located deep in the ligation pocket (W611, Y623, Y689 and Y798) have the most severe effect. W611A reduces the activity to less than 1%, and both step-2 and step-3 ligation show undetectable activity. The Hen1-N with the first amino acid deleted (ΔM1) also shows significant reduction in activity, which is supported by the fact that the alignment of Hen1 members indicated strict conservation of precise N terminus. Besides those five mutations mentioned above, the remaining ones within the first group are R63A, K614A and K824A, which are all positive residues and may be involved in guiding the RNA substrate. However, we cannot rule out the possibility that those mutations may affect the global structure by destructing the important interaction.

The second group of mutations contains four mutants L61A, Y77A, H194A and Q622A (Figure 2.14D), which have only modest effects on the enzymatic activity. Generally, they all slightly reduced activity, but the reduction is less than one half. From structural view, they are all located relatively distant from the active site, so those mutations may not affect substrate binding or enzymatic activity that much.

The remaining 6 mutants including L74A, L618A, E862A, D867A, R869A and ΔL870 (Figure 2.14D) show increased enzymatic activity. They displayed more than 50% increase of activity, especially for the step-2 ligation. Most of those residues are from C-additional module of Pnkp-C, and contribute to the formation of ligation pocket. We speculate that the mutations may help to alter the ligation pocket both in size and charge and make it more accessible to the RNA substrates. Since we don't really know the *in vivo* substrates for Pnkp/Hen1 RNA repair

system, and the Pnkp-C-ΔL/Hen1-N may show different features compared to full-length proteins, it's hard to explain the biological significance of those strictly conserved residues.

4. Figures and Tables

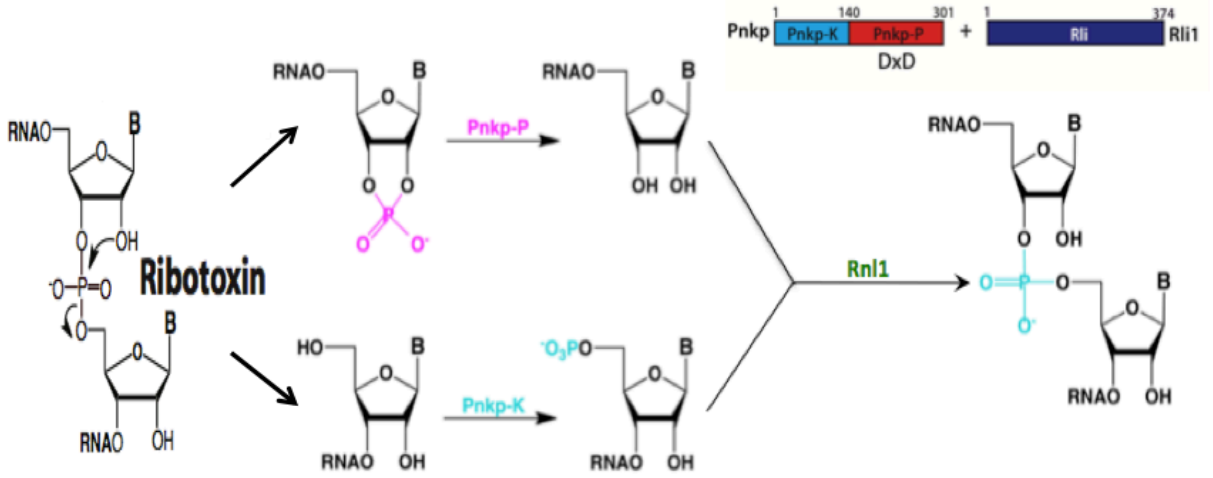

Figure 2.1 *T4*Pnkp/Rli carry out RNA repair. Ribotoxin catalyzes transesterification reaction and generates two halves of RNA. *T4*Pnkp processes the broken ends of RNA: phosphatase domain (Pnkp-P) removes the 2', 3'-cyclic phosphate while kinase domain (Pnkp-K) phosphorylates the 5'-OH. *T4*Rli ligates the processed RNA to restore the original form.

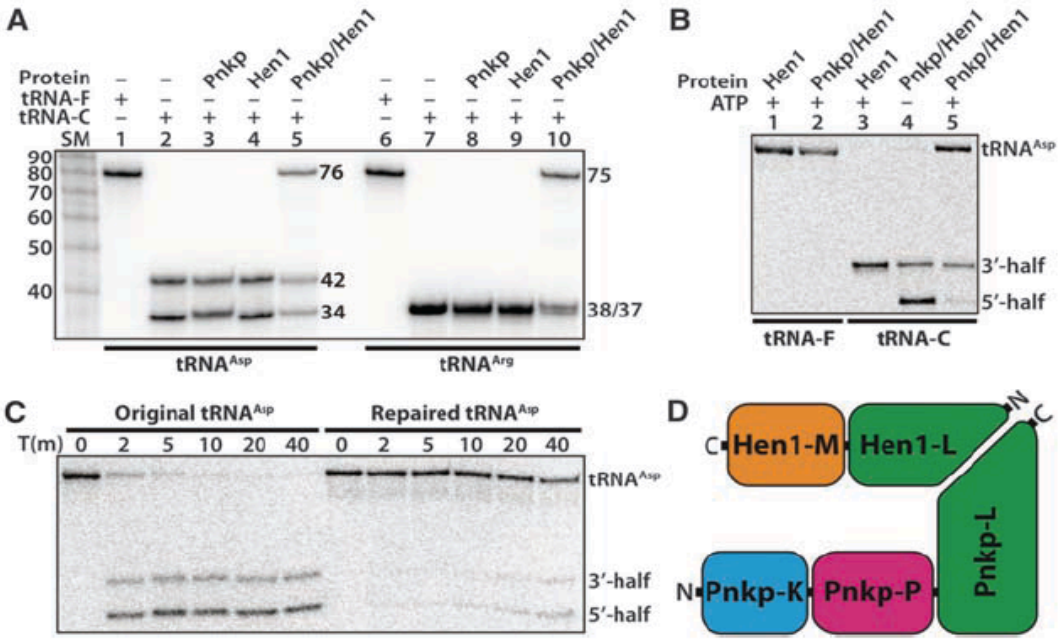

Figure 2.2 Pnkp/Hen1 repairs the broken RNA and enhances the product (34). (A) Pnkp/Hen1 repair cleaved tRNA (tRNA-F is full length tRNA, tRNA-C is cleaved tRNA); (B) Methylation of tRNAAsp by ^{14}C-AdoMet. Only the substrate with free 3'-end can be methylated by Hen1, and Pnkp/Hen1 can repair the methylated tRNA in presence of ATP; (C) The repaired tRNA^{Asp} is resistant to the same cleavge; (D) Schematic representation of Pnkp/Hen1 complex.

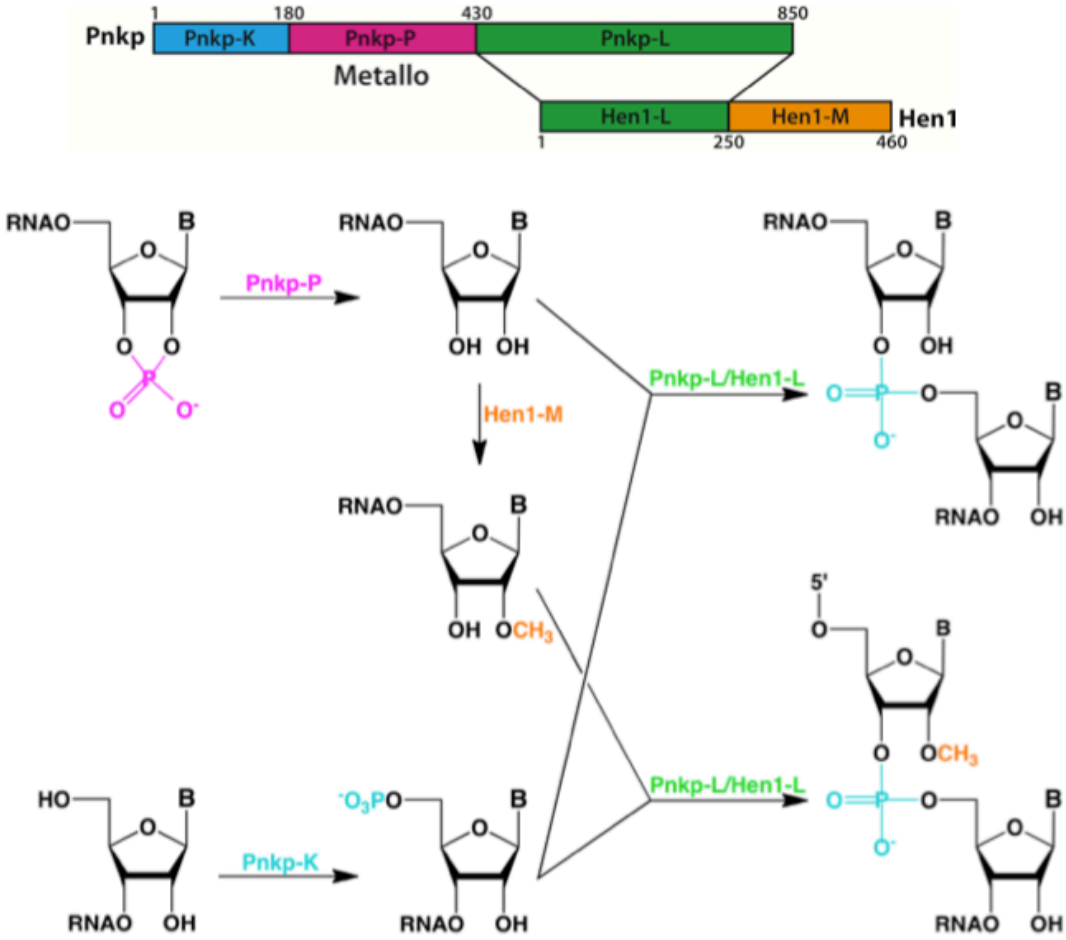

Figure 2.3 RNA repair carried out by bacterial Pnkp/Hen1. Compared to *T4*Pnkp/Rli (Figure 2.2), the two ends of broken RNA are also processed by phosphatase (Pnkp-P) and kinase domain (Pnkp-K), respectively. Unlike*T4*Pnkp/Rli, the methyltransferase domain of bacterial Hen1 (Hen1-M) also carries out extra 2'-O-methylation at the 3'-end. Moreover, the ligation steps require doamins from both Pnkp and Hen1 (Pnkp-L/Hen1-L) in bacteria.

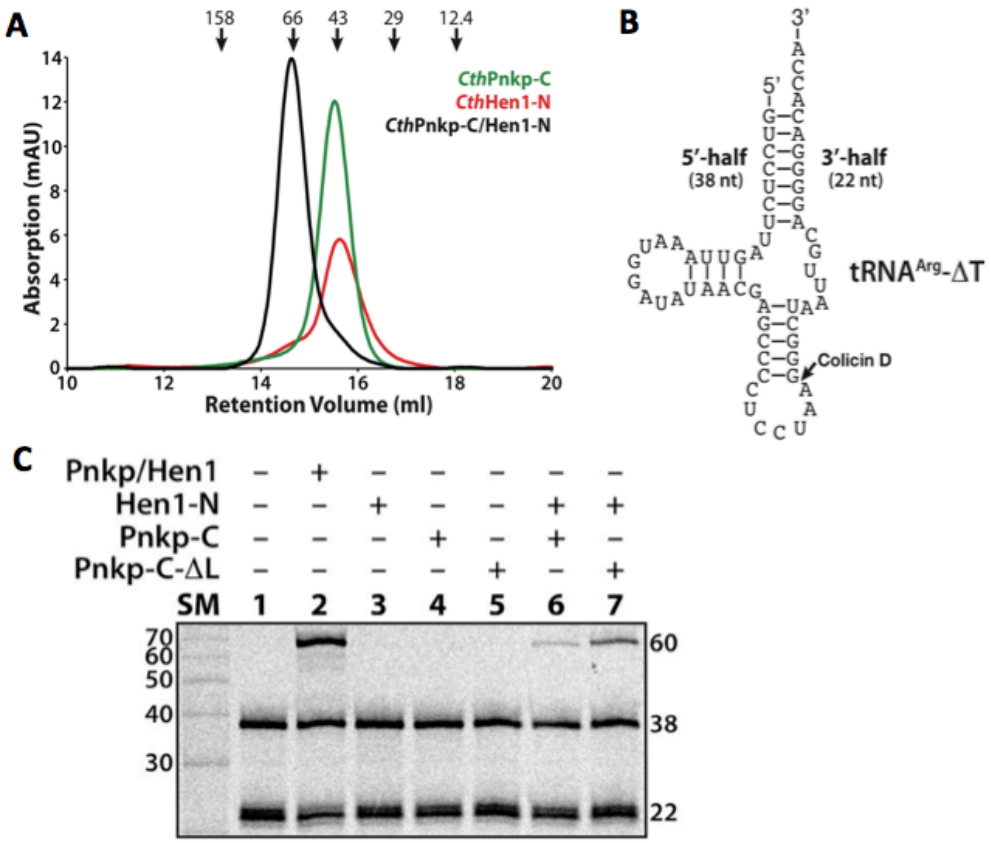

Figure 2.4 A stable *Cth*Pnkp-C/Hen1-N heterodimer formed in vitro is an active RNA ligase. (A) Size-exclusion chromatography of the purified CthPnkp-C (green), *Cth*Hen1-N (red), and their equal molar mixture (black). Positions of retention volumes of the protein standards are marked with arrows and labeled with their molecular weights (kDa). (B) $tRNA^{Arg}$ – ΔT was used in the in the following experiment, and its clevage site is indicated by the arrow. (C) Ligation of the processed two halves of the cleaved $tRNA^{Arg}$-ΔT by different combinations of *Cth*Hen1-N, *Cth*Pnkp-C, and *Cth*Pnkp-C-ΔT. The ligation by the full length *Cth*Pnkp/Hen1 complex (lane 2) was used as the positive control. SM, size marker.

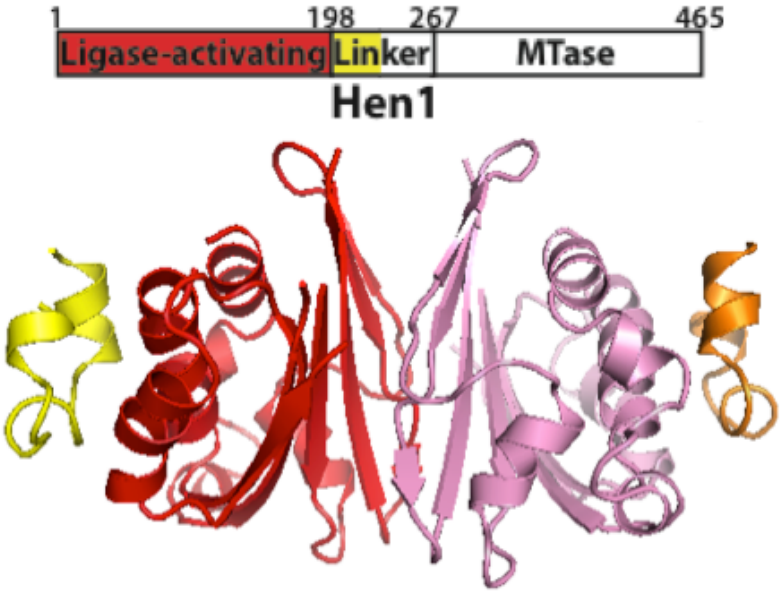

Figure 2.5 Overall structure of *Cth*Hen1-N homodimmer. Each monomer is seven-stranded β-sheet stacked with two layers of α-helices. The red part (pink part in the other monomer) is assigned as ligase-activating domain (interacting with Pnkp-C), while the yellow part (orange part in the other monomer) linked to the missing methyltransferase domain of Hen1. Two monomers interact through their β-sheet domains.

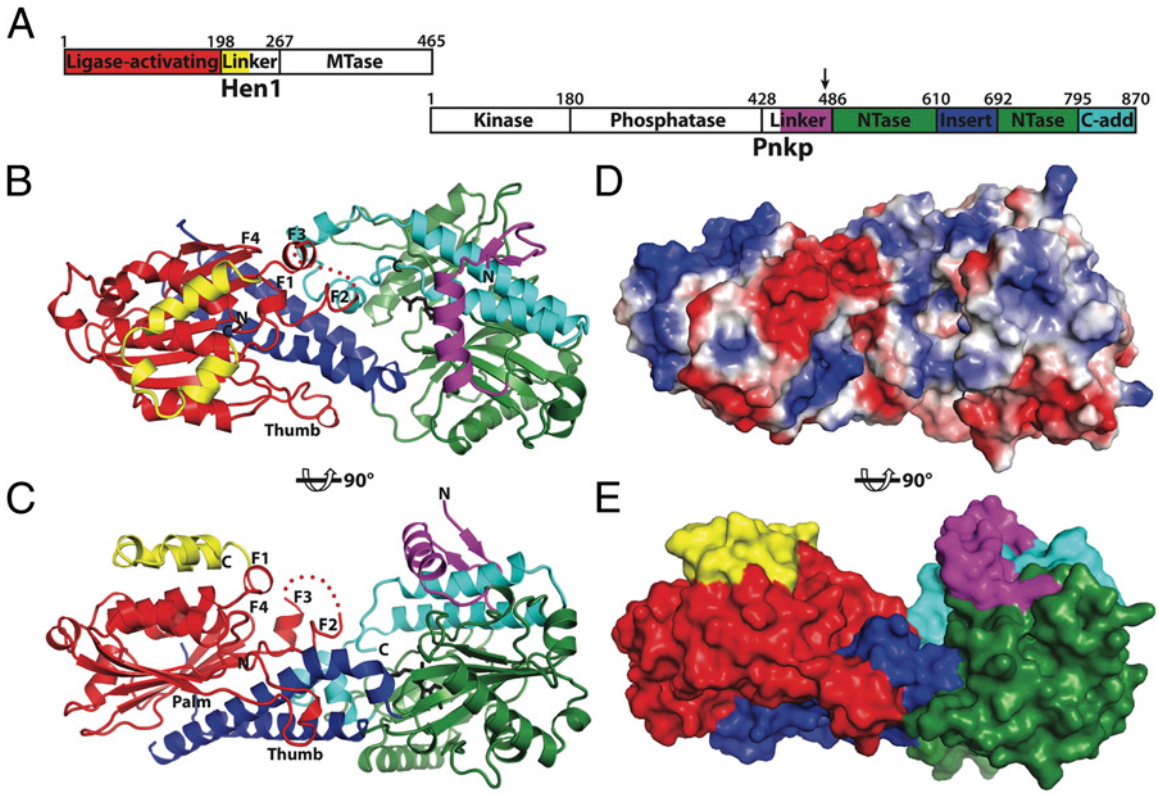

Figure 2.6 Overall structure of the *Cth*Pnkp-C/Hen1-N/AMP ternary complex. (A) Schematic representation of a refined assignment of domains in *Cth*Hen1 and *Cth*Pnkp. The regions in color correspond to the *Cth*Pnkp-C/Hen1-N heterodimer. An arrow denotes the start site of a shorter version of *Cth*Pnkp-C (termed CthPnkp-C-ΔL by its essential lack of the linker region), which we used for functional analysis. C-add, C-addition; Insert, insertion; MTase, methyltransferase; NTase, nucleotidyltransferase. (B and C) Ribbon representation of the top (B) and side (C) views of the structure of the ternary complex. The structure is in the same color codes as shown in A. The covalent AMP is in stick and colored black to provide the location of the active site. A disordered loop in *Cth*Hen1-N is depicted with dots. The ligase-activating domain of *Cth*Hen1-N is depicted as a left hand, with the palm, the thumb, and four fingers (F1–F4) labeled. (D) Surface of the *Cth*Pnkp- C/Hen1-N heterodimer in the same orientation as B and colored for electrostatic potential (blue, basic; white, neutral; red, acidic). (E) Surface of the *Cth*Pnkp-C/ Hen1-N heterodimer oriented and colored the same as C to highlight the grip of the insertion module of *Cth*Pnkp-C (blue) by the ligase-activating domain of *Cth*Hen1-N (red).

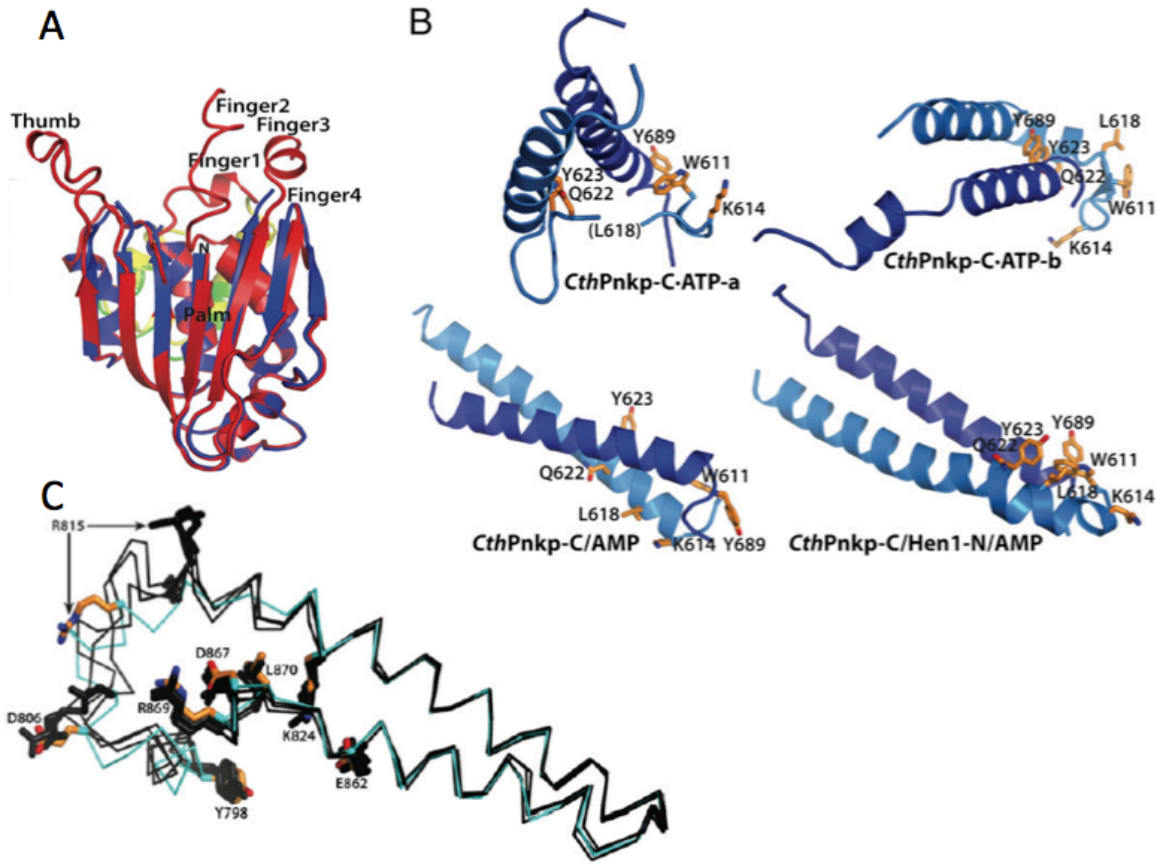

Figure 2.7 Structural changes from the *Cth*Hen1-N homodimer and the *Cth*Pnkp-C monomer to the *Cth*Pnkp-C/Hen1-N heterodimer. (A) Superposition of the structure of one copy of *Cth*Hen1-N from *Cth*Hen1-N homodimer (blue and green) and from *Cth*Pnkp-C/Hen1-N heterodimer (red and yellow). The rmsd of the two structures is 0.52 Å. The structure of the ligase-activating domain is depicted as a left hand, with the β-sheet as the palm and five loops as the thumb and fingers. The loops are only structured when Hen1-N binds to Pnkp-C. (B) Ribbon representation of the structures of the insertion module from the *Cth*Pnkp-C·ATP (PDB: 3TY5), the *Cth*Pnkp-C/AMP (PDB: 3TY9), and the *Cth*Pnkp-C/Hen1-N/AMP complexes. The structure is in the same orientation as the one shown in Figure 2.6. Two helices of the structure are colored with different shades of blue to highlight their relative positions. The side chains of six conserved residues near the base of the insertion module are in stick and colored orange, with heteroatoms colored individually (nitrogen in blue and oxygen in red). (C) Cα superposition of the C-addition module of *Cth*Pnkp-C from the *Cth*Pnkp-C/Hen1-N heterodimer (cyan) and from *Cth*Pnkp-C·ATP (black) and *Cth*Pnkp-C/AMP (black). The side chains of eight conserved residues are highlighted in stick.

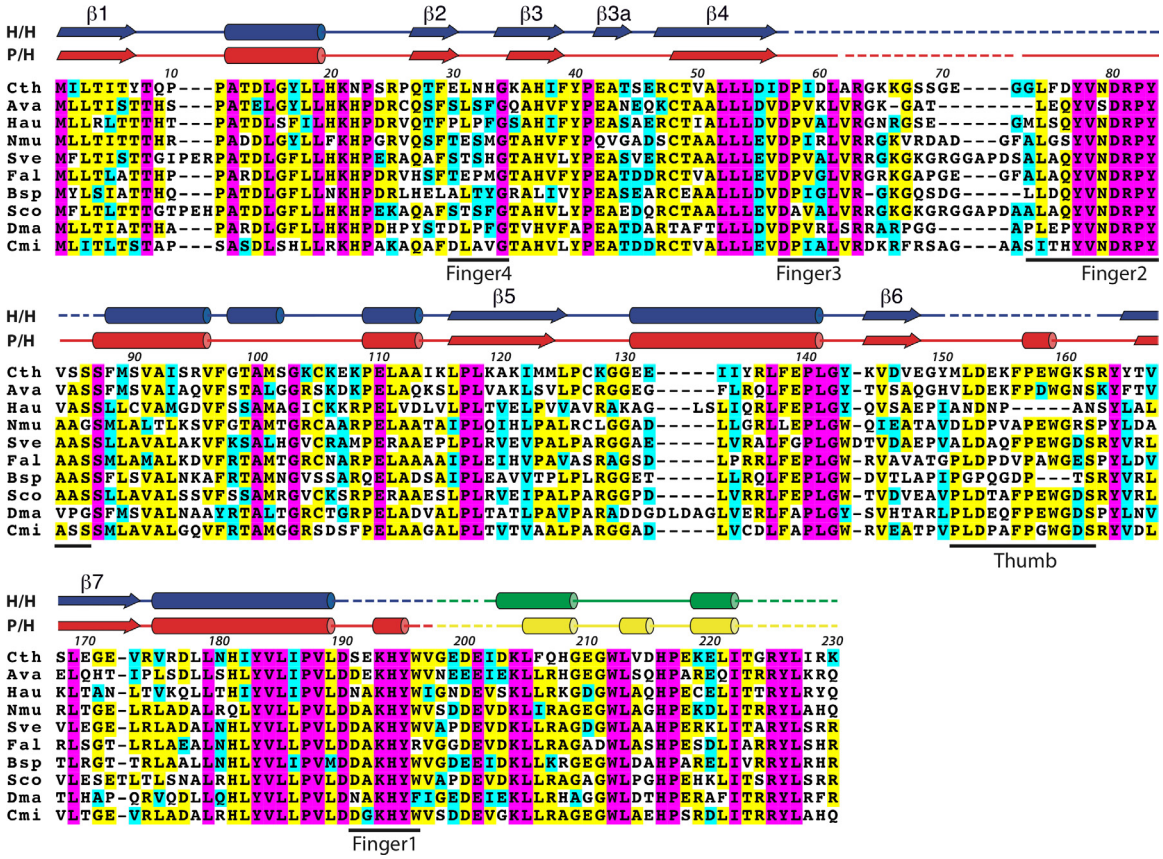

Fig 2.8 Conservation of Hen1-N in bacteria. The conserved residues are boxed in color, with completely conserved residues in magenta, identical residues in yellow, and similar residues in cyan. Residue number over the alignment corresponds to *Cth*Hen1-N. The secondary structure of *Cth*Hen1-N is depicted above the primary sequence, with α-helices highlighted as cylinders, β-strands as arrows, loops as solid lines, and disordered residues as dotted lines. The structure of *Cth*Hen1-N is portrayed as a left hand (Figure 2.7 A), and the regions corresponding to the thumb and fingers are underlined and labeled. *Ava, Anabaena variabilis; Bsp, Bradyrhizobium sp. BTAi1; Cmi, Clavibacter michiganensis; Cth, C. thermocellum; Dma, Deinococcus maricopensis; Fal, Frankia alni; Hau, Herpetosiphon aurantiacus; H/H, CthHen1-N homodimer; Nmu, Nakamurella multipartita*; P/H, *Cth*Pnkp-C/Hen1-N heterodimer; *Sco, Streptomyces coelicolor; Sve, Streptomyces venezuelae*.

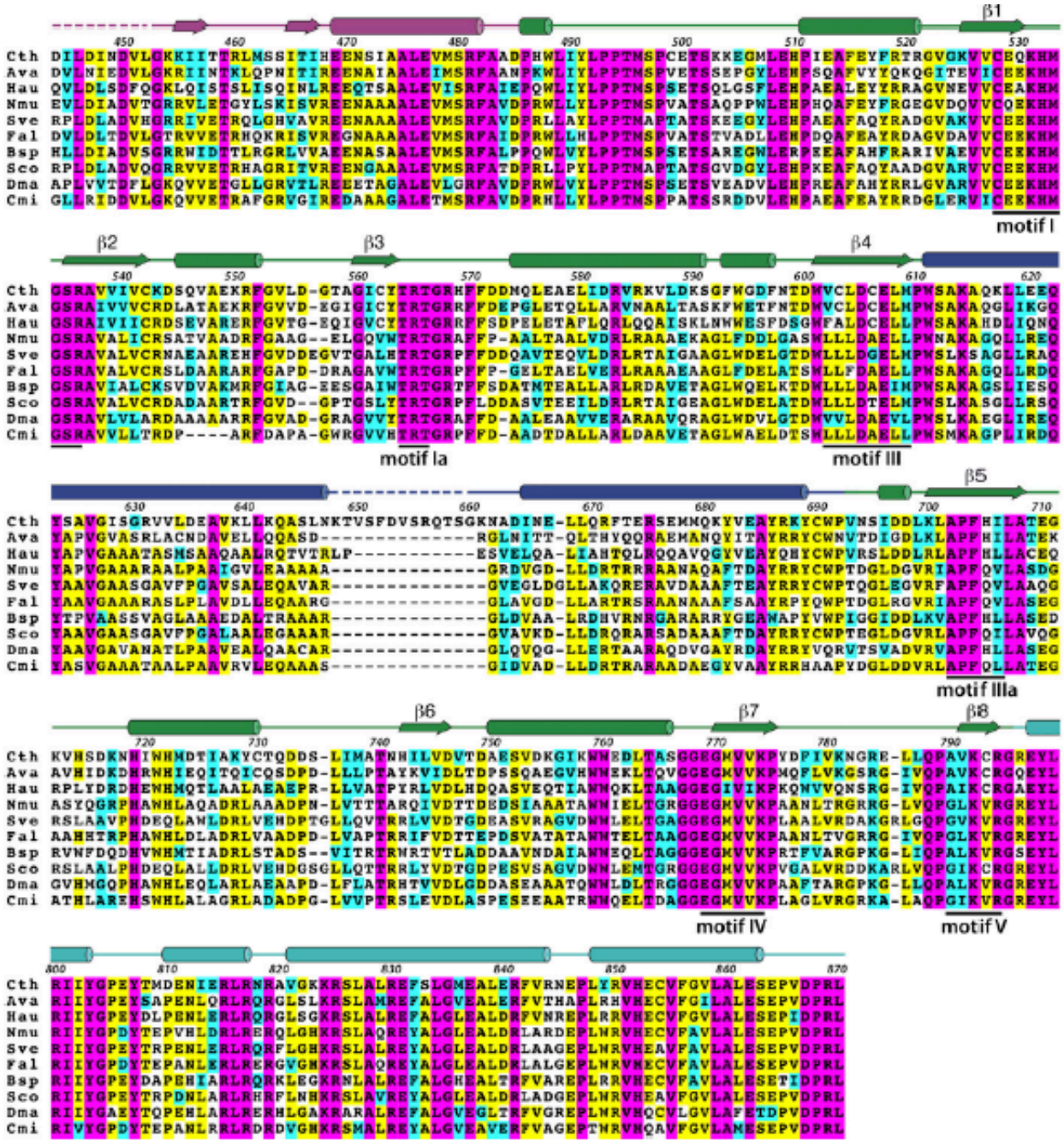

Figure 2.9 Conservation of the C-terminal half of Pnkp in bacteria. The organisms and their orders are the same as those for Hen1-N shown in Figure 2.8. Residue number over the alignment corresponds to *Cth*Pnkp-C. The NTase motifs conserved in DNA ligases, RNA ligases, and RNA-capping enzymes are underlined and labeled

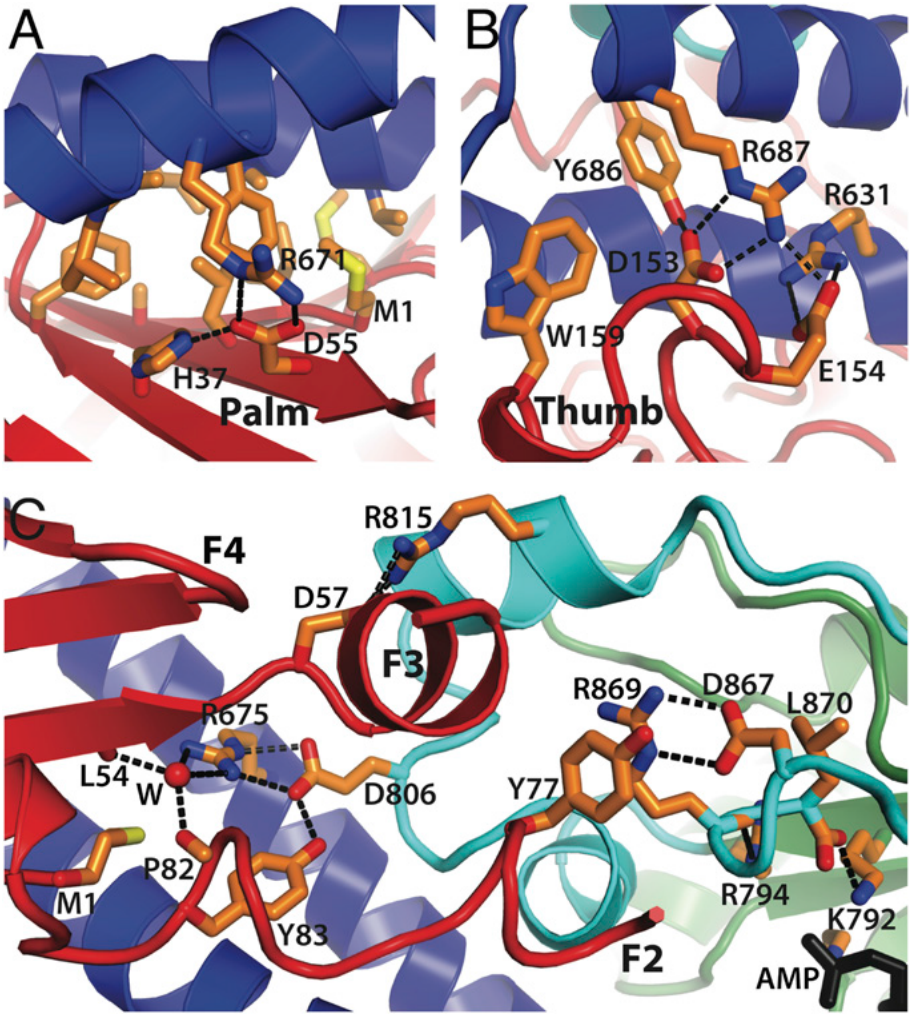

Figure 2.10 Details of the heterodimer interface. (A) The interface between the palm of the ligase-activating domain and the insertion module. Cα-chains of the structure are represented and colored the same as in Figure 2.6A. The side chains are in stick and colored orange except heteroatoms, which are colored individually (nitrogen in blue, oxygen in red, and sulfur in yellow). Hydrogen bonds and salt bridges are depicted with black dashed lines. Many hydrophobic residues involved in interaction between the palm and the insertion module are not labeled for clarity. (B) Details of the interface between the thumb of the ligase-activating domain and the insertion module. (C) Details of the interface of the three-way interaction among finger2 and finger3 of the ligase-activating domain, the insertion module, and the C-addition module. A water molecule is depicted in sphere and colored red. The covalent AMP is partially shown to provide the location of the active site.

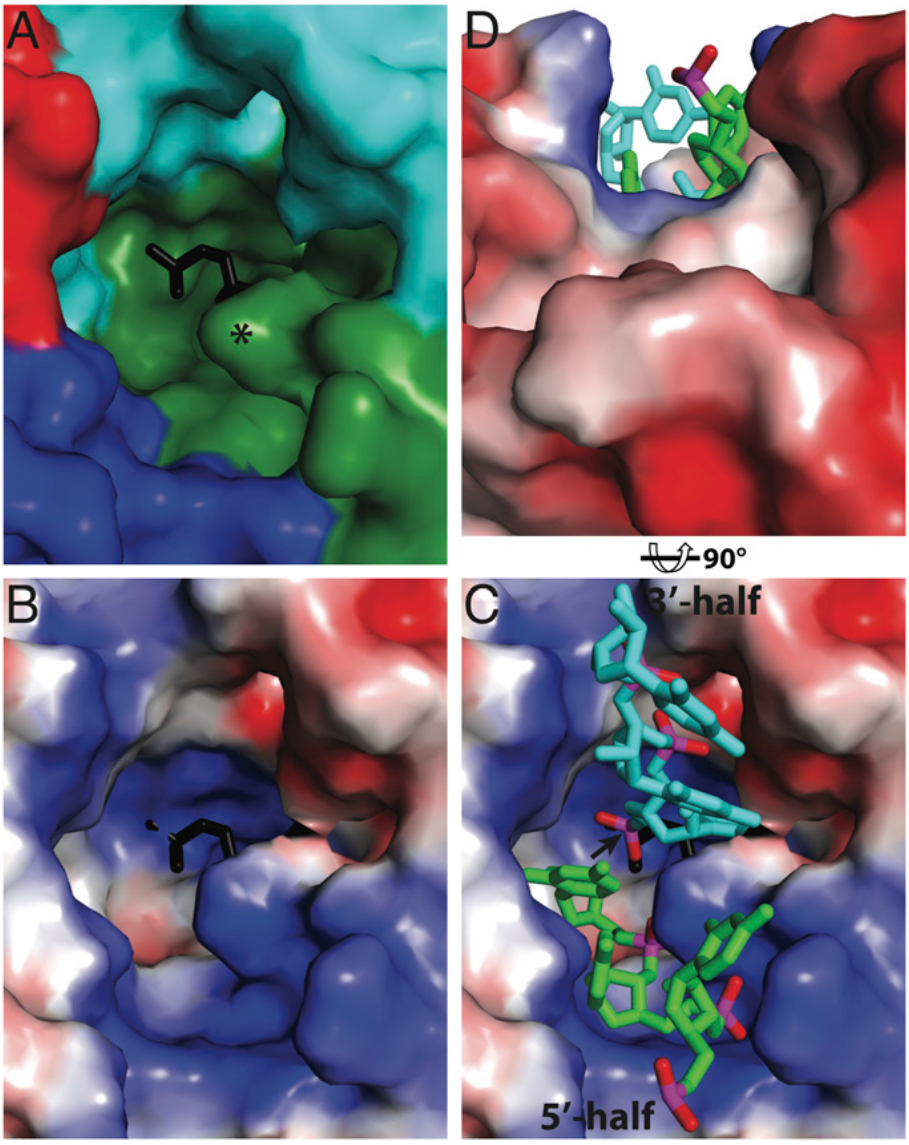

Figure 2.11 Ligation pocket in *Cth*Pnkp-C/Hen1-N. (A) Surface of the ligation pocket in the same color codes as in Figure 2.6A. The covalent AMP is in stick and colored black. The side chain of R565 that partially divides the pocket is marked with an asterisk. The structure of the linker region of *Cth*Pnkp-C is omitted for clarity. (B) The same surface as Figure 2.11A but colored for electrostatic potential as shown in Figure 2.6D. (C) Modeling of two strands of RNA (three nucleotides each) into the ligation pocket. The modeled 5'-half and 3'-half RNA substrate are in stick and colored green and cyan, respectively. Phosphate groups are colored magenta (phosphorus) and red (oxygen) to highlight their locations. The potential phosphodiester bond formed between the 5'-phosphate and the 3'-OH is marked with a black arrow. (D) Side view of the ligation pocket shows that the third nucleotides of both strands of the modeled RNA are out of the ligation pocket.

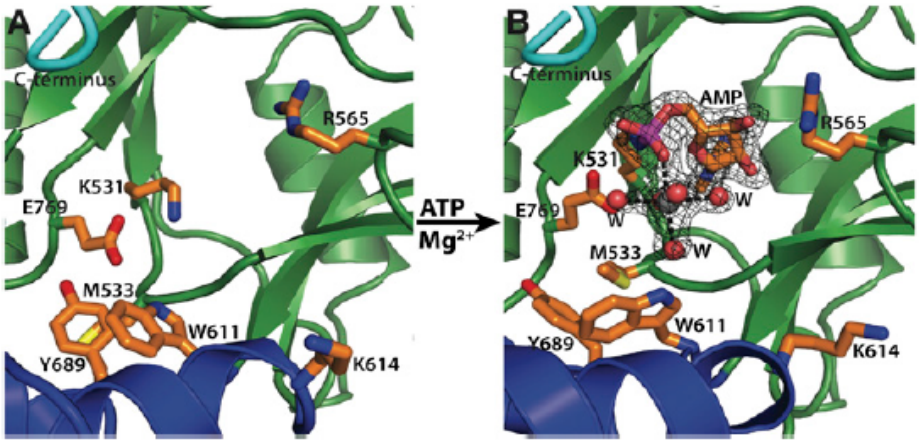

Figure 2.12 Structural changes of *Cth*Pnkp-C/Hen1-N heterodimer accompanying with the adenylation. The structural details of the active site of the *Cth*Pnkp-C/Hen1 heterodimer without (A) or with (B) ATP and Mg^{2+} soaked before data collection. The structures are colored and depicted in a manner similar to those shown in Figure 2.10. Only the AMP and the side chains of residues with significant conformational changes are shown in stick. An Mg^{2+} ion and five coordinated water molecules are depicted in spheres and colored gray and red, respectively. Mg^{2+} and a water molecule are not labeled for clarity. The simulated annealing omit map covering AMP, $Mg^{2+,}$ and five water molecules is contoured at 1.5 σ.

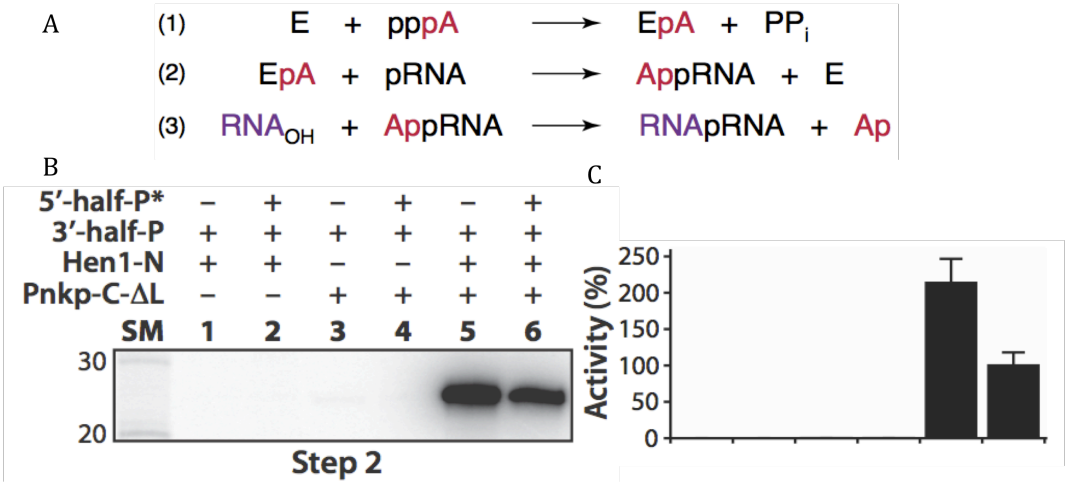

Figure 2.13 Step 2 ligation assay of heterodimer. (A) The three ligation steps are conserved in RNA ligase, DNA ligase and RNA capping enzyme; (B) Analysis of step 2 ligation carried out by *Cth*Hen1-N, *Cth*Pnkp-C-ΔL, or a combination of both using either one-stranded RNA (3'-half-P) or two-stranded RNA (5'-half-P*/3'-half-P) as the substrate. The presence of 2', 3'-cyclic phosphate at the 3'-end of 5'-half-P* allows step 2 to occur but blocks step 3 due to the lack of 3'-OH. (C) Data in the bar graph are represented as means ± SD of three separate experiments.

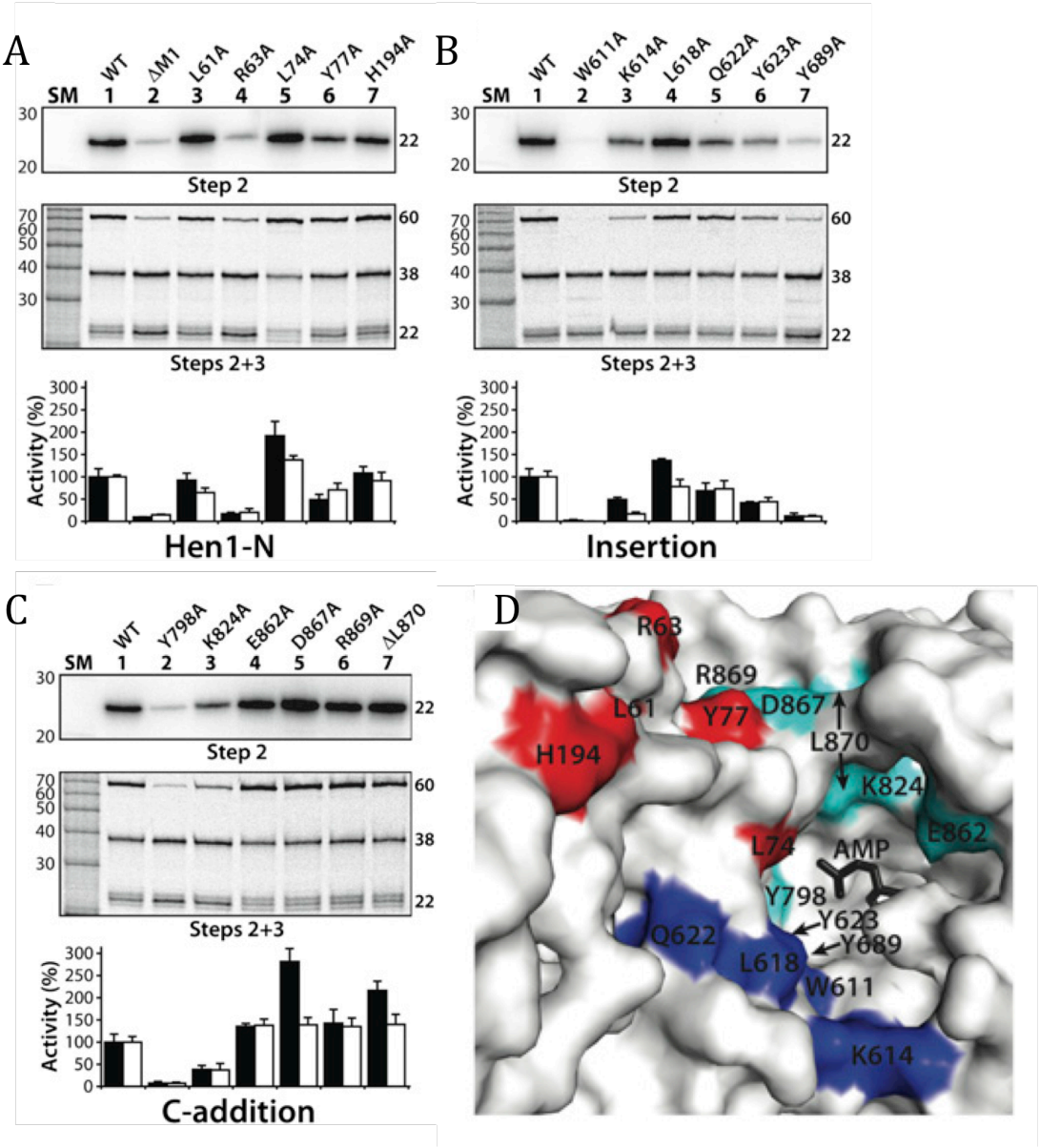

Figure 2.14 Functional assays of step 2 and step 3 with the wild type and mutant enzymes. (A-C) DPAGE analyses of step 2 (Top) and steps 2 + 3 (Middle) carried out by the wild type and mutant enzymes. The RNA substrate for step 2 is 5'-half-P*/3'-half-P, and the substrate for steps 2 + 3 is the ^{33}P-internally labeled cleaved tRNA^{Arg}-ΔT. The reactions were quantified based on the amount of radioactivity incorporated into 3'-half-P (step 2) or the percentage of the cleaved tRNA^{Arg}-ΔT to be repaired (steps 2 + 3), followed by normalization to the activity of the wild-type enzyme, which is defined as 100%. The black bar represents the relative activity for step 2, and the white bar representsthe relative activity for steps 2 + 3. (B) Surface of an expanded view of the ligation pocket shows the locations of 18 residues (except M1 of *Cth*Hen1-N, which is distant from theligation pocket) selected for mutation and deletion analyses. Residues are colored the same as those shown in Figure 2.6A.

Table 2.1 Data collection and refinement statistics for crystals in chapter 2

Data collection	*Cth*Pnkp-C/Hen1-N/AMP	*Cth*Pnkp-C/Hen1-N	*Cth*Hen1-N
Space Group	P1	P1	$P6_5$
a, b, c (Å)	59.6, 59.8, 101.5	59.7, 60.0, 101.7	68.5, 68.5, 165.7
α, β, γ (°)	81.4, 87.8, 88.8	81.2, 87.6, 88.3	90.0, 90.0, 120.0
Resolution (Å)	50.0–2.4	50.0–2.6	50.0–2.3
Wavelength (Å)	0.9787	0.9787	0.9787
Redundancy	3.7	3.8	7.5
Completeness (%)*	89.7 (79.7)	98.0 (94.2)	98.7 (100)
Average I/σ	14.8 (2.9)	11.8 (2.5)	14.9 (2.8)
R_{merge} (%)	5.2 (19.7)	5.6 (23.0)	6.1 (20.4)
Refinement			
Resolution (Å)	34.6–2.4	42.4–2.6	40.4–2.3
No. reflections	49,244	41,067	19,215
R_{work}/R_{free} (%)	16.8/22.2	18.5/24.5	20.4/25.7
No. atoms	10,530	10,046	2,796
Protein	10,054	9,886	2,734
Ligand/ion	46	6	N/A
Water	430	154	62
Average B factor, $Å^2$	36.9	48.6	56.3
rmsd values			
Bond lengths (Å)	0.08	0.010	0.010
Bond angles (°)	1.08	1.28	1.38
Ramachandran plot			
Favored (%)	97.3	96.5	97.2
Outliners (%)	0.3	0.7	0.3

*Highest resolution shell is shown in parenthesis.

CHAPTER 3: BIOCHEMICAL AND STRUCTURAL STUDIES OF BACTERIAL RNA REPAIR COMPLEX PNKP1/RNL/HEN1

1. Introduction

There is always conflict for a limited resource of nutrition between organisms, microbes in particular. Some organisms employ a variety of toxins to eliminate competitors, while others use molecular defense system to neutralize toxins for survival. Protein synthesis may be the biggest target for toxins due to its conservation and essential role in living organisms. Most ribotoxins employ a transesterification mechanism to cleave essential RNAs (rRNA, tRNA) involved in protein synthesis and generate a hydroxyl group at 5'-end and 2', 3'- cyclic phosphate at the 3'-end (9-23). As described in chapter 2 (Figure 2.1), bacterial Pnkp/Hen1 system is able to carry out the repair of damaged RNA together with 2'-O-methylation at the repair conjunction, resulting in a product resistant to the repeated damages from the same toxin (Figure 2.2, 2.3) (34, 35).

Because methylated RNA repair product is superior, there is great incentive for the Pnkp/Hen1 repair complex to have evolved to maximize the production of methylated product. In theory, this can be achieved if the opportunity and efficiency of methylation can be ensured for this complex. Our previous study as well as that of Shuman's group solves the opportunity issue (31, 58). Shuman and coworkers found that the ligase domain of Pnkp is disabled by the flexible insertion domain that blocks the ligase active site, while our study revealed that N-terminal domain of Hen1 activated ligase activity of Pnkp by locking the insertion domain of Pnkp into a conformation that allows access to the ligation pocket (Figure 2.6, 2.11).

To address the methylation efficiency issue, it requires knowledge of the relative locations of active sites of phosphatase, methyltransferase, and ligase, all of which must be

traversed by the 3'-end of the damaged RNA with that particular order to produce the methylated RNA repair product. However, we failed to produce diffraction quality crystals of the Pnkp1/Hen1 heterotetramer. Instead, we discovered a new bacteria RNA repair complex consisting of three proteins Pnkp1, Rnl and Hen1 (Figure 3.1), which form a heterohexamer in vitro. The new RNA repair complex and Pnkp/Hen1 share some features despite some difference in sequence and catalytic mechanism. The structural and biochemical study on the new RNA repair complex provides molecular insight into efficient 2'-O-methylation during repair, which may hold true for the Pnkp/Hen1 RNA repair complex due to their similarity.

2. Method and Materials

2.1. Cloning, overexpression and purification of recombinant protein

Genes encoding Pnkp1, Rnl and Hen1 were amplified from the *Capnocytophaga gingivalis* genomic DNA purchased from ATCC and inserted into the pETDuet-1 vector. The encoding plasmids were transformed into *E. coli* BL-21(DE3) strain individually, and the proteins were expressed at 18 ºC for 20 hours after induction with 0.5mM IPTG. Cells were harvested by centrifugation and stored at -80 °C.

Thawed cell pellets were resuspended in lysis buffer (20 mM Tris-HCl, pH 8.0, 10mM NaCl, 2% glycerol, and 1mM DTT) and lysed by French press. The cell lysate was centrifuged and the proteins were purified from the supernatant by FPLC system.

Pnkp1 was purified to homogeneity by using DEAE ion exchange, Heparin affinity, Mono Q ion exchange and Superdex 200 size exclusion chromatography. Rnl was purified the same as Pnkp1 except that Mono Q ion exchange step was omitted. Hen1 was purified similarly as Pnkp1 except that Heparin affinity chromatography was skipped. All purified proteins were

stored in the gel filtration buffer (10 mM HEPES, pH 7.0, 200 mM NaCl, and 1 mM DTT) for further study.

To produce selenomethione-incorporated Pnkp1, *E. coli Rosetta* strain was used for protein expression and methionine pathway inhibition was used for cell growth. The protein was purified the same way as the wild-type Pnkp1 described above.

2.2. *In vitro* reconstitution

The purified recombinant proteins were analyzed individually using size exclusion chromatography. To assess protein-protein interaction, two different proteins were combined in equimolar and incubated at 4°C for an hour before size exclusion chromatography. To assemble the entire RNA repair complex, all three proteins were mixed in equimolar concentrations, incubated at 4 °C for an hour, and analyzed by size exclusion chromatography.

To carry out RNA repair, ^{33}P-internally radiolabeled and colicin D-cleaved $tRNA^{Arg}$ was prepared as described previously. In a 20 μl reaction volume, the cleaved tRNA (4 μM) was incubated with different combinations of proteins (1 μM each) in the repair buffer (25 mM Tris-HCl, pH 8.0, 50 mM KCl, 2.5 mM $MgCl_2$, 0.5 mM $MnCl_2$, 0.05 mM EDTA, 5 mM DTT, and 2.5% glycerol) in the presence of 0.2 mM ATP and 0.05 mM AdoMet at 37 °C for 45 min. After the reaction, the sample was processed with phenol extraction followed by ethanol precipitation to recover RNA. RNA was dissolved in DPAGE loading buffer, heated at 95 °C for 3 min to denaturation, and analyzed by 15% DPAGE. The radioactivity of the repaired and unrepaired tRNAs was visualized using a PhosphorImager system (Molecular Dynamics).

2.3 Kinetic study of enzymatic activities

In this part of study, $tRNA^{Arg}$-ΔT was used as the substrate, which can be cleaved by colicin D and generate 5'-half (38mer) and 3'-half (22mer)

To compare the kinase activity for *Cgi*Pnkp1 and *Cgi*Pnkp1/Rnl/Hen1 heterohexamer, synthetic 22mer RNA (corresponding to 3'-half of $tRNA^{Arg}$-ΔT) was purchased from DNA IDT. In the 30μl reaction volume, 10 μM 22mer RNA was incubated with *Cgi*Pnkp1 or *Cgi*Pnkp1/Rnl/Hen1 (0.025 μM each) in the repair buffer at 37 °C. After 0.2 mM ATP (containing 1 μCi of ^{33}P-alaph-ATP) was added into the reaction, samples were taken out at the following time point: 0, 5, 10, 30, 60 minutes and processed with phenol extraction followed by ethanol precipitation. The recovered RNA was dissolved in DPAGE loading buffer, heated at 95 °C for 3 minutes and analyzed by 15% DPAGE. The radioactivity was visualized and quantified by a PhosphorImager system (Molecular Dynamics).

To prepare the substrate for dephosphorylation, the cleaved $tRNA^{Arg}$-ΔT (38mer and 22mer) was repaired by *Cgi*Pnkp1/Rnl/Hen1 heterohexamer in the repair buffer (0.2 mM ATP was added) in the presence of 10 μCi of ^{33}P-alaph-ATP. The repaired product was gel purified and subject to the cleavage by colincin D. The cleaved product 38mer, that contained radioactive cyclic phosphate at the 3'-end, was purified and labeled as 38P*. In the 30 μl reaction volume, 1μM 38P* RNA was incubated with *Cgi*Pnkp1 or *Cgi*Pnkp1/Rnl/Hen1 (0.25 μM each) in the repair buffer at 37 °C. 5 μl of reaction solution was taken out at the following time point: 0, 5, 10, 30, 60 minutes and quenched by adding DPAGE loading buffer. All the samples were heated at 95 °C for 3 minutes and loaded into 15% DPAGE. The analysis of radioactivity was the same as described before.

For the ligation activity assay, the internal radiolabelled 38mer RNA (labeled 38*) was

prepared through 4ml in vitro transcription reaction with 20 μCi ^{33}P-alaph-GTP added. 22mer RNA with 5'-phosphate (22P) was purchased from DNA IDT. In a 30 μl reaction volume, 1μM annealed 38*/22P and 0.25 μM proteins *Cgi*Rnl/Hen1 were incubated in the repair buffer (0.2 mM ATP was added) without or with 0.05 mM AdoMet at 37 °C. The following sampling and analysis were the same as dephosphorylation assay.

In order to check methylation efficiency, the reactions described in the ligation activity assay were scaled up by 100 times. The repaired products (61mer) were gel purified and labeled as 61*Sam-, 61*Sam+. 10 μM of 61*Sam- or 61*Sam+ was dissolved in 1X repair buffer, heated up to 95 °C for 3 minutes and quickly cooled down on ice for 5 minutes. A reaction mixture of 30 μl scale containing the cleavage buffer (25 mM HEPES, pH 7.0, 25 mM NaCl, 5 mM $MgCl_2$, 0.05 mg/ml BSA), 1 μM quickly annealed RNA (61*Sam- or 61*Sam+) and 1μM colicin D was incubated at 25°C. The following sampling and analysis were the same as before.

2.4 Crystallization, data collection, and structural determination.

The purified Pnkp1/Rnl/Hen1 heterohexamer was concentrated to 7mg/ml, and mixed with a reservoir solution containing 8% PEG6000, 0.2 M NaCl, 15 mM $MgCl_2$, and 100 mM MES (pH 6.2). Full size crystals of the Pnkp1/Rnl/Hen1 heterohexamer grew through hanging drop vapor diffusion method at 4 °C in two weeks. Crystals were soaked in a series of cryoprotecting solutions containing all the components of the reservoir solution supplemented with increased percentage of glycerol to a final concentration of 30%. For the data set that produced the structure of the Pnkp1/Rnl/Hen1 heterohexamer in complex with ATP, 5mM ATP was also included in crystal soaking. The cryo-protected crystals were mounted in nylon loops and flash-frozen in liquid nitrogen. Data was collected at 21-ID beamline at the Advanced

Photon Source (APS) and processed by HKL200035. Phase for the structure of the hexamer was determined based on a SAD data from a crystal of the heterohexamer with the selenium-containing Pnkp1. The initial model was built automatically with the Phenix software. The remaining model was built using Coot program, followed by refinement using Phenix program (68,69). Many rounds of model building, followed by refinement, resulted in a final model of the Pnkp1/Rnl/Hen1 heterohexamer with Rwork and Rfree of 17.3 % and 23.7 %, respectively.

2.5 Metagenomic DNA assay

Amino acid sequences of *Cgi*Hen1 of the Pnkp1/Rnl/Hen1 RNA repair complex were employed for BLASTP search against metagenomic data released by the Human Microbiome Project (https://img.jgi.doe.gov/cgibin/imgm_hmp/main.cgi). The E-value for the search was 1e-50. Because of significant difference between Hen1 from the Pnkp1/Rnl/Hen1 RNA repair complex and the one from the Pnkp/Hen1 system, the E-value of 1e-50 ensures that a positive result from a search is an indication of the presence of the Pnkp1/Rnl/Hen1 RNA repair system, not the Pnkp/Hen1 system, in the dataset used for the search.

The entire metagenomic analyses consist of three stages. First, BLASTP search using *Cgi*Hen1 was carried out against datasets of each of the five groups (Airway, Gastrointestinal tract, Oral, Skin, and Urogenital tract). Second, the search was carried out against datasets of each of the sub-groups of human mouth (Supragingival plague, Subgingival plague, Throat, Tongue dorsum, Attached/Keratinized gingival, Palatine tonsils, Buccal mocosa, and Saliva). Third, BLASTP search using the sequences of Hen1 from six human-hosted bacteria was carried out one at the time against the datasets of Supragingival plague, Subgingival plague and Throat. In addition to the E-value of 1e-50, the positive result of the search also required the sequence to

be at least 97% identical in order to distinguish different Hen1 from these six bacterial species.

3. Results and Discussions

3.1 Discovery of the new RNA repair system in bacteria

The Pnkp/Hen1 RNA repair complex is found in ~250 bacterial species, and both proteins are highly conserved and exhibit ~40-60% sequence identities. We used the sequence of *Ava*Hen1 (34) to perform comprehensive BLASTP search and discovered a distinct Hen1gene existing in ten bacterial species. This newly discovered Hen1 is modestly homologous (20%) to *Ava*Hen1, and there is no gene encoding Pnkp within the same operon. Instead, an operon that is far away from this hen1was found to contain two genes encoding proteins that may be involved in RNA repair: one encodes a protein that is equivalent to the ligase domain of Pnkp with the modest sequence identities (~20%), and the other encodes a protein homologous to *T4*Pnkp (~30% identity) (Figure 3.1). Because these two proteins are functionally equivalent to the Pnkp in bacterial Pnkp/Hen1 system, we hypothesize that these two proteins, together with the newly found Hen1, constitute a new RNA repair system. Since this new system possesses elements of both T4 Pnkp/Rnl and bacterial Pnkp/Hen1 repair systems, it can be regarded as a hybrid RNA repair system. To distinguish the components of new RNA repair system from the Pnkp/Hen1, we tentatively assign the ligase homolog as bacterial Rnl (RNA ligase), and the T4 Pnkp homolog as bacterial Pnkp1.

3.2 Reconstitution of the Pnkp1/Rnl/Hen1 RNA repair complex *in vitro*

To test the hypothesis that bacterial Pnkp1, Rnl, and Hen1 may constitute a new RNA repair system, we cloned the genes encoding these three proteins from *Capnocytophaga gingivalis* (*Cgi*) and purified the three recombinant proteins individually (Figure 3.2c).

We first carried out chromatographic analyses with the purified proteins to probe the interactions among them. Size exclusion chromotogrphy revealed that Pnkp1 formed a homodimer in solution, while the other two proteins exist as monomers. Analyses of pair-wise mixture in equimolar indicated that Pnkp1 and Rnl formed heterotetramer, Rnl and Hen1 formed a heterodimer, but Pnkp1 and Hen1 did not interact. When all three proteins were present in the equimolar mixture, a single species was formed. Judged by the elution volume, the three proteins formed a Pnkp1/Rnl/Hen1 heterohexamer (Figure 3.3b).

To assess the RNA repair capability, we employed a ribotoxin-cleaved $tRNA^{Arg}$ as the substrate (Figure 3.4A), which was previously used for in vitro reconstitution of Pnkp/Hen1 repair complex (34, 41). The damaged tRNA could be only repaired in the presence of all three proteins (Figure 3.4B). We also used cleaved rRNA as substrate and observed the similar result (Figure 3.5). Therefore, Pnkp1/Rnl/Hen1 heterohexamer is likely the functional unit of the new RNA repair system in vivo. Because this heteroheamer possesses two identical copies of each activity, a total of 8 active sites exist in the complex, which is same as in bacterial Pnkp/Hen1 heterotetramer.

3.3 Kinetic study of enzymatic activities

To confirm four enzymatic activities of the Pnkp1/Rnl/Hen1 heterohexamer, we performed additional biochemical assays aiming at individual enzymatic activity (Figure 3.6). For these assays, we employed $tRNA^{Arg}$-dT, which allowed us to purify both 5'-half and 3'-half

of RNAs generated by colicin D (cleavage of full-length $tRNA^{Arg}$ produces two RNAs with 37 and 38 nucleotides, which cannot be separated). Moreover, Cleaved $tRNA^{Arg}$-dT has previously shown to be an effective substrate for RNA repair by Pnkp/Hen1 heterotetramer (41).

Among the enzymatic reactions we have assayed, RNA 5'-phosphorylation reaction performed by the kinase domain is the fastest (Fig. 3.6a). When compared to Pnkp1 homodimer, formation of Pnkp1/Rnl/Hen1 heterohexamer results in a modest rate enhancement in 5'-phosphorylation (Fig. 3.6a, compare the curve marked with square to the one marked with cycle compare the two time curves).

3'-dephosphoyrlation, carried out by the phosphatase domain of Pnkp1, is slower compared to 5'-phosphorylation (Figure 3.6b). As in the case of kinase reaction, formation of Pnkp1/Rnl/Hen1 heterohexamer has a modest positive effect on the dephosphorylation reaction as well (Figure 3.6b, compare the curve marked with square to the one marked with cycle).

To assess the effect of 2'-methylation on RNA ligation, the RNA substrates with processed ends were subjected to ligation reaction carried out by the Rnl/Hen1 heterodimer in the absence or presence of S-adenosylmethionine (SAM) (Figure 3.6c). The presence of SAM greatly increases the rate and efficiency of RNA ligation (Figure 3.6c, compared the curve marked with square to the one marked with cycle). Judged by the initial reaction rate (for example, the data points of 5 min reaction), the rate enhancement is at least 10-fold. The mechanism of this rate enhancement is unknown and requires further investigation. But it is possible that the ligation pocket of Rnl has evolved to accommodate 2'-methylated substrate better.

When the RNA repair products shown in Figure 3.6c were purified and subjected to colicin D cleavage, the repaired RNA produced in the presence of SAM showed more resistance

to cut when compared to the one without SAM (Figure 3.6d, compare the curve marked with square to the one marked with cycle), indicating that 2'-O-methylation by Hen1 during RNA repair results in immunity of the repaired RNA. The incomplete cleavage of the repaired RNA without 2'-methylation might be due to the partial misfolding of $tRNA^{Arg}$-dT (Figure 3.6d, the curve marked with cycle).

3.4 Overall structure of Pnkp1/Rnl/Hen1 heterohexamer

We crystallized the Pnkp1/Rnl/Hen1 heterohexamer and solved the structure at 3.3Å resolution (Figure 3.7). The structure confirmed all the protein-protein interactions observed in chromatographic experiment, but also revealed the unexpected physical interaction between Pnkp1 and Hen1. In the structure, each protein connects with the other two different kinds of proteins, resulting in Pnkp1/Rnl/Hen1 heterotrimer adopting a ring structure. Due to the dimerization of Pnkp1, the entire hexamer structure is like two rings fused at the Pnkp1 homodimer interface (Figure 3.8a, c). The side view of the structure displays C2 symmetry, which indicates that the RNA substrates approach this complex from two opposite direction (Figure 3.8b, d).

The Pnkp1 homodimer is formed through the interactions between the same enzymatic domains, resulting in the kinase and phosphatase dimer modules (Figure 3.10a). The kinase module and phosphatae modules essentially do not interact with each other, with the exception of the N-terminus and the linker region connecting kinase and phosphatase. The relative orientation of kinase and phosphatase modules is locked at 130° in the hexamer structure (Figure 3.10b), however, we can speculate that the two modules can adopt flexible relative orientation when there is no spatial restriction from other interacting proteins, Rnl and Hen1. The formation of

homodimer leads to ~ 6400Å^2 buried area. Each kinase domain of Pnkp1 homodimer interacts with one Rnl, resulting in ~600Å^2 buried area. Each Rnl interacts with one copy of Hen1, leading to ~3200 Å^2 buried area. However, one copy of Hen1 interacts with two copies of Pnkp1 through their kinase and phosphatase domain, respectively, and ~700Å^2 surface area is buried due to this three-way interaction (Figure 3.7a).

3.5 Detailed interactions and comparison

The crystal structure reveals that how the hexamer is formed through protein-protein interaction. Moreover, the Dali search based on each active domain helps to identify the enzymatic function and mechanism (67).

3.5.1 Formation of Pnkp1 homodimer

The formation of kinase module is via the interaction of two kinase domains with antiparallel orientations. The interactions is mainly hydrophobic, and contributed by the side chains of V39, F46, M49, R63, M67 and L75 from both kinase domains. Moreover, the interaction is further enhanced by the several hydrogen bonds, including one at the center (S71 from both domains) and two at each end (K66-N78 and R63-T80) (Figure 3.9a). Dali search revealed that the fold of the kinase domain and the mode of the dimerization are similar to some *T4*Pnkp structures (67, 70-72); even the majority of amino acids involved in dimerization are not conserved (Figure 3.11a, b).

Unlike kinase module, the formation of phosphatase module is resulted from the interaction between two parallel phosphatase domains. The formation of a hydrophobic core, provided by the side chains of V211, V214, M217, Y223, F284, L286, F305 and V307, is also the major driving force for the dimerization (Figure 3.10b). The interaction is further enhanced

by two hydrogen bonds at the bottom of the interface (pair Y307-D294). As in the case of kinase module, Dali search showed that the fold of phosphatase domain and the mode of phosphatase dimerization are similar to some *T4*Pnkp structures (67, 73-76). The residues involved in dimerization of phosphatase are highly conserved, together with that the mutations of those conserved residues ablate the phosphatase function in *T4*Pnkp, indicates that the dimerization of phosphatase in *Cgi*Pnkp1 is essential for the dephosphorylation activity (Figure 3.11a, c).

3.5.2 Interaction between Rnl and Pnkp1

Each Rnl only interact with one kinase domain of Pnkp1 homodimer, and the interaction is mostly electrostatic (Figure 3.9c). The main interactions are salt bridge formed by the charged residues from both Pnkp1 (R29, R33, D36 and K110) and Rnl (K4, E369 and D379). An additional hydrogen bond formed by the side chain of R33 in Pnkp1 and N375 in Rnl is also involved.

3.5.3 Recognition of Rnl by Hen1

The C-terminal methyltransferase structures of Hen1 from both Pnkp/Hen1 and Pnkp1/Rnl/Hen1 systems are well aligned (Figure 3.12a) (35). As was previously observed in the structure of ligase module in Pnkp/Hen1 RNA repair complex, the recognition of Rnl by Hen1 is presumably the molecular basis of ligase activation. The interaction mainly occurs at two locations: first, the β sheet of the N-teminal ligase-activating domain of Hen1 recognizes the insertion domain of Rnl via an extensive hydrogen-bonding network; second, an extended loop of the N-terminal ligase activating domain of Hen1 reaches over the dimer interface and interacts with the C-terminal part of Rnl via formation of several hydrogen bonds (Figure 3.9d).

Even the interaction modes are similar for Pnkp-C/Hen1-N and Rnl/Hen1, there is substantial differences in both structure and detailed interactions, especially for the insertion domains and ligase-activating domains from these two systems (Figure 3.12 b-e) (58). Since the ligase-activating domain was shown to activate the ligase activity by locking the conformation of insertion domain, the pronounced differences of these two parts imply an early evolutionary divergence of these two repair systems.

3.5.4 Interaction between Hen1 and Pnkp1 homodimer

The interaction between Hen1 and Pnkp1 was not observed in the chromatographic analysis of binary mixture of these two proteins. However, Hen1 seemed to interact with both copies of Pnkp1 in the structure (Figure 3.9e). This three-way interaction is mainly hydrogen bonds network occurring among the 20 aa peptides segment of Hen1 (S289, R311, D315 and Y318), one loop region from phosphatase domain of one Pnkp1 (D257 and R262), N-terminal segment from kinase domain of the other copy of Pnkp1 (S2, N5 and F10).

The interaction between Hen1 and Pnkp1 homodimer leads to ~700 $Å^2$ buried area, but the interaction is quite weak in general. We speculate that the interaction is not stable in the absence of Rnl, which helps to bring Pnkp1 homodimer and Hen1 close enough.

3.6 Implication of the mechanism

To provide insight into the mechanism of RNA repair by Pnkp1/Rnl/Hen1 heterohexamer, we soaked crystals with ATP before data collection, which results in the crystal structure with cofactors occupying all active sites (Figure 3.13). The Pnkp1/Rnl/Hen1

heterohexamer contains two sets of active sites, and each set contain four different types of enzymatic functions required for one complete round of RNA repair process.

3.6.1 Pnkp1 is a bifunctional enzyme

The N-terminal domain of Pnkp1 is a kinase, which belongs to P-loop phosphotransferase superfamily. The main chain amide and the side chain of serine in the P-loop coordinate the β phosphate of ATP (Figure 3.14a) (72). We docked a polynucleotide substrate in this active site, and proposed that Arg47 might help to stabilize the substrate phosphate backbone, while Asp44 could activate 5'-OH of substrate through the abstraction of its proton. Activated 5'-OH is readily to attack the γ phosphate of ATP with the help from Lys24, which stabilizes transition state of phosphate.

The C-terminal domain of Pnkp1 is phosphatase, which belongs to acylphosphatase superfamily. The signature motif for this superfamily is DxDxT, which coordinates metal and water molecule to facilitate the removal of 2', 3'- cyclic phosphate at 3'-end of RNA substrate (Figure 3.14b). According to the study of *T4*Pnkp, phosphatase convert a 2', 3'-cyclic phosphate to 2'-OH and 3'-OH through sequential diesterase and monoesterase reactions (Figure 3.14c) (75). In the process, Asp183 is required for the phosphate removal through covalent aspartyl-phosphate intermediate, and Asp185 would donate a proton to the leaving groups, ribose O3.

3.6.2 Hen1 contains 2'-O-methyltransferase

The C-terminal domain of Hen1 is S-adenosylmethionine (SAM) -dependent methyltransferase (MTase), which carries out 2'-O-methylation at the 3'-terminal nucleotide of RNA. As mentioned before, this part is quite conserved in Pnkp/Hen1 and Pnkp1/Rnl/Hen1

system, which is also confirmed by the structure alignment of *Cth*Hen1-C and *Cgi*Hen1-C. The core fold of MTase consists of seven-stranded β-sheet flanked by helices on both sides (Figure 3.12a). In our structure, we could see relatively low density of S-adenosylhomocysteine (SAH) (Figure 3.13c), which is demethylation product of SAM. We modeled a strand of RNA and a metal ion in the active site (Figure 3.15a). The side chains of conserved residues glutamate and histidine, together with 2' -OH and 3' -OH group of RNA substrate could coordinate the metal ion. In the previous study of Hen1, it has been shown that the metal ion, usually Mn^{2+}, could precisely position the 2'-OH toward the methyl group of SAM and facilitate the methylation (Figure 3.15b) (77). After methylation of RNA substrate, the metal ion would be released due to the lack of coordination. We could not see the density of metal ion in our structure for the same reason.

3.6.3 RNA ligase

The Rnl belongs to covalent NTase superfamily, and it alone is not able to carry out RNA ligation even though it contains all the conserved motifs of RNA ligase. We could see the density of ATP at the active site (Figure 3.13d). Lys61 is proposed to attack α phosphate of ATP to form lys-AMP covalent adduct, which is first step of RNA ligation. Guided by the ligand, we model two strands of RNA in the active site (Figure 3.15c). The enzyme could activate 5'-end of one strand of RNA by transferring AMP, and then the 3'-OH of the other strand of RNA is readily to attack the activated 5'-end to form bond (Figure 3.15d) (59).

3.6.4 Proposed reaction pathway

To understand why the RNA repair complex requires two copies of each protein, the two units of the Pnkp1/Rnl/Hen1 heterotrimer must first be defined. According to the previous discussion about gene arrangement, we know that Rnl and Pnkp1 are co-translated in *C. gingivalis*, while Hen1 is synthesized separately. We speculate that Rnl and Pnkp1 should form a complex first (heterotetramer) (Figure 3.3b), and then each Rnl binds to on copy of Hen1 for the hexamer formation. Thus the copy of Rnl and Pnkp1 those make physical contact, together with the copy of Hen1 that interacts with Rnl should be assigned as the same unit. The labeling of two units of Pnkp1/Rnl/Hen1 heterotrimer in the figure takes into consideration of these in vivo sequential events (Figure 3.8a).

Guided by the location of the cofactors in all the active sites, and aided by the published structural homologs of Pnkp1 and Hen1 in complex with nucleic acids as well as the RNA docking model in the ligase module Pnkp-C/Hen1-N (58, 72, 73, 77), we manually docked short single-stranded RNAs (ssRNA) in the active sites located in the inner rim of the same ring structure (Figure 3.16a). The docking model shows that, while the reacting ends of RNA approach the active sites, the opposite ends point to the vacant space at the center of the ring. Moreover, the vacant space is only big enough to adopt a double-stranded RNA, which may indicate how the double-stranded part of broken RNA fit into the complex and help to bring the two broken ends to the active sites (Figure 3.17). We proposed that the damaged RNA approaches the four active sites from the one side of the ring structure (Figure 3.8d, 3.16b). The 5'-end of the damaged RNA enters the kinase site for the phosphorylation, while 3'-end of the damaged RNA goes into the phosphatase site for dephosphorylation. After phosphorylation, the 5'-end of RNA directly transfer to the ligase site, however, the dephosphoryalted 3'-end of RNA makes a detour to the methyltransferase site for 2'-O-methylation before goes into the ligase site.

RNA ligation is carried out as long as both processed ends of damaged RNA are present at the ligase active site. The pathway for RNA repair carried out by the Pnkp1/Rnl/Hen1 is depicted in Figure 3.16b, which results in the repaired RNA product with 2'-O-methylation at the repair conjunction.

According to relative location of each active site, we proposed that a single processive RNA repair event is performed by the four active sites located at the inner rim of each ring (Figure 3.16b). Utilizing four active sites contributed from both rings for a repair event would require the RNA substrate transfer among the active sites that are far away from each other, and the constant association and dissociation between protein and RNA would reduce the efficiency of 2'-O-methylation which is the responsible to the immunity for the further damages. The structure illustrates that both copies of Pnkp1/Rnl/Hen1 heterotrimer make contribution to the each set of active sites, and confirms that the hexamer is functional unit for RNA repair.

3.6.5 Comparison of two RNA repair systems in bacteria

The study of Pnkp1/Rnl/Hen1 RNA repair system may provide insight into how Pnkp/Hen1 heterotetramer work. These two RNA repair complexes are evolutionary related according to sequence blast search. We are not able to get the structure of entire Pnkp/Hen1 complex, but already solved some functional modules, such as *Cth*Pnkp-C/Hen1-N (ligase module) and *Cth*Hen1-C (methyltransferase). The structural comparison of those parts to corresponding modules in Pnkp1/Rnl/Hen1 revealed that these two systems are very similar, especially in the core structures and interaction mode. We expect that the Pnkp/Hen1 heterotetramer to be similar in structure to the portion of the Pnkp-C/Hen1 (Figure 3.12).

However, the manner how the two units of Pnkp-C/Hen1 are brought together by Pnkp-N to form heterotetramer is unclear. First, even *Cth*Pnkp-N is functionally similar to *Cgi*Pnkp1, both containing kinase and phosphatase activities, but the phosphatase parts of these two systems belong to completely different superfamilies; second, in Pnkp1/Rnl/Hen1 structure, ligase module makes contact with kinase domain while in Pnkp gene, ligase module is physically connected to phosphatase domain (Figure 3.1). Therefore, the structures corresponding to the kinase and phosphatase domains of these two repair systems are expected to be different.

Nevertheless, Pnkp/Hen1 heterotetramer and Pnkp1/Rnl/Hen1 heterohexamer contains the same type and number of active sites, and both of them are able to carry out the same biochemical reactions. The two-fused-ring overall structure, and the RNA repair mechanism revealed by the Pnkp1/Rnl/Hen1 heterohexamer structure are likely to be conserved in Pnkp/Hen1 heterotetramer. Despite extensive in vitro studies of the Pnkp/Hen1 as well as Pnkp1/Rnl/Hen1 RNA repair complex, the in vivo substrates still remain unknown, which may be due to that the RNA damage and repair occurring in the wilderness is hard to be repeated in the laboratory setting. Pervious Pnkp-C/Hen1-N structure as well as this Pnkp1/Rnl/Hen1 heterohexamer structure revealed that all the active sites are readily accessible to a variety of damaged RNAs. This is consistent with the pervious in vitro biochemical studies that Pnkp/Hen1 RNA repair complex exhibit broad substrate specificity (41). Therefore, instead of repairing a particular RNA damaged by a particular ribotoxin, both bacterial RNA repair complex might be generic RNA repair systems that repair a variety of damaged RNA in vivo as long as the broken ends can be brought into active sites.

3.7 Metagenomic analysis

Despite having little knowledge about the in vivo biological function of Pnkp1/Rnl/Hen1 RNA repair complex, we were able to identify the bacterial species that possess the Pnkp1/Rnl/Hen1 complex. Among of ten species, six were isolated from human, one was isolated from dog, and the rest are isolated from environmental samples with unknown resources. The data released by the Human microbiome project (HMP) shows that among five locations of human where the sample are collected, the bacteria possessing the RNA repair complex are only present in human mouth, especially in gingival plagues (Figure 3.18a, b) (78-81). Further analysis of individual bacterium aginest the data indicated that *Capnocytophaga gingivalis*, *Capnocytophaga sputigena*, and *Capnocytophaga sp. taxon* 326 are the most abundant (Figure 3.18c). Among ~5000 bacterial species with known genome, only a subset of *Capnocytophaga* species that mainly live in gingival plagues of human month possess the Pnkp1/Rnl/Hen1 RNA repair complex. Especially, Both *C. gingivalis* and *C. sputigena* have been implicated in periodontal and dental diseases. Even we don't know the clear reason for this correlation, we speculate that the unique RNA repair carried out by Pnkp1/Rnl/Hen1 complex may provide these bacteria with a heightened ability to survive in human mouth. The study related to Pnkp1/Rnl/Hen1 may provide the potential application for human health.

4. Figures and Tables

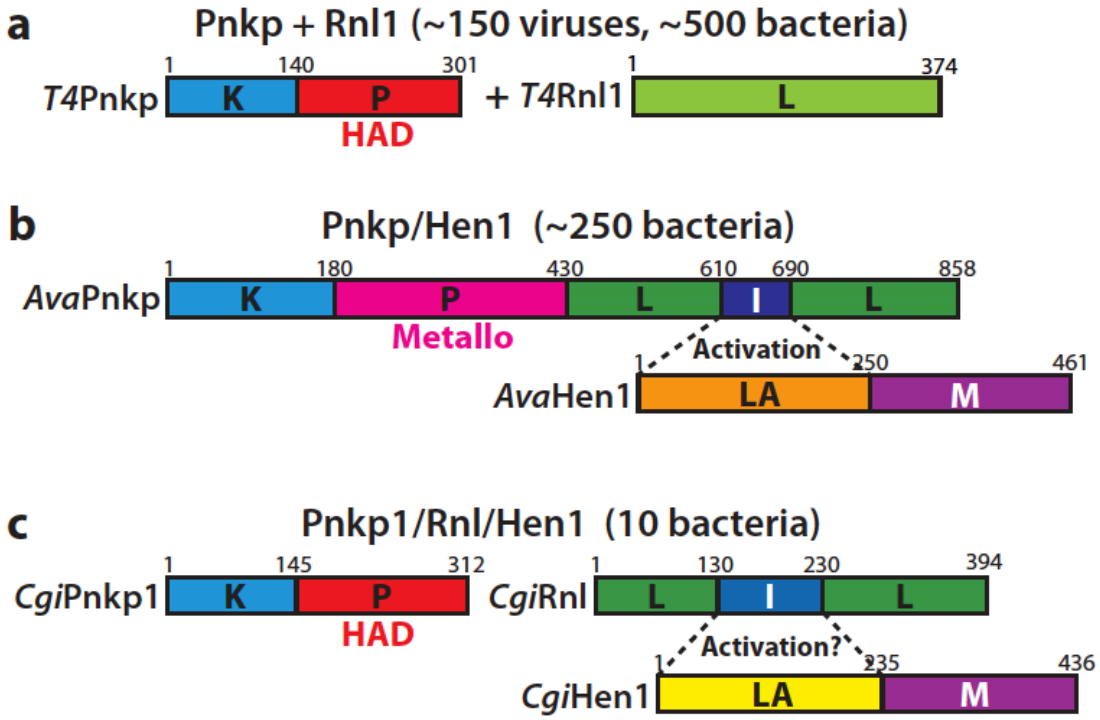

Figure 3.1 Overview of RNA repair systems. (a) Schematic view of the two proteins from bacteriophage *T4* in RNA repair. The similar RNA repairsystem is present in approximately 150 other viruses and 500 bacterial species. (b) Schematic view of the Pnkp/Hen1 RNA repair complex from Anabaena variabilis (Ava). The Pnkp/Hen1 RNA repair system can be found in about 250 bacterial species of different phyla. (c) The the newly discovered Pnkp1/Hen1/Rnl RNA repair complex from *Capnocytophaga gingivalis* (*Cgi*). This system is currently only found in ten bacteria belonging to the Flavobacteriaceae family of the Bacteroidetes phylum. K, kinase; P, phosphatase; M, methyltransferase; I, insertion domain; LA, ligase-activating domain.

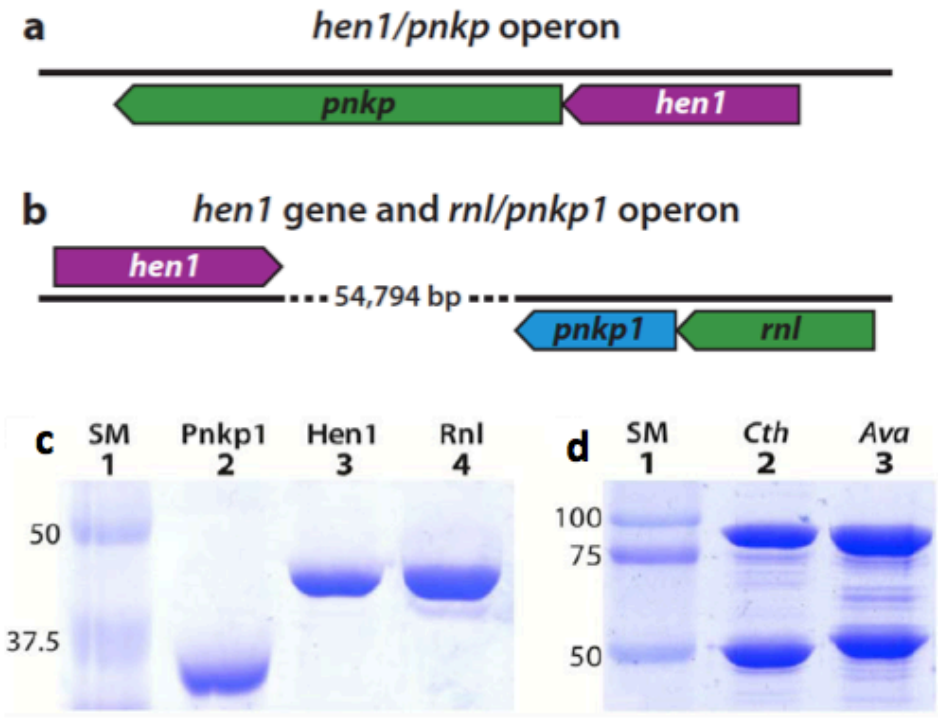

Figure 3.2 Gene organization of the two representative bacterial RNA repair Systems (*Ava*, *Cgi*) depicted in Figure 3.1. (a) The genes encoding two proteins of the *Ava*Pnkp/Hen1 RNA repair complex are present in the same operon. The two proteins are cotranslational, with Hen1 synthesized first. (b) The newly discovered *Cgi*Pnkp1/Rnl/Hen1 RNA repair complex has the gene encoding Hen1 separated from the ones encoding Rnl and Pnkp1. *Cgi*Rnl and *Cgi*Pnkp1 are co-translational, with Rnl synthesized first. 54,794 base pairs (bp) of DNA separate the gene encoding Hen1 and the ones encoding Rnl and Pnkp1. (c) SDS gel analyses of the purified proteins constituting the Pnkp1/Hen1/Rnl from *Cgi*. (c) Purified protein Pnkp/Hen1 complex from *Ava* and *Cth*.

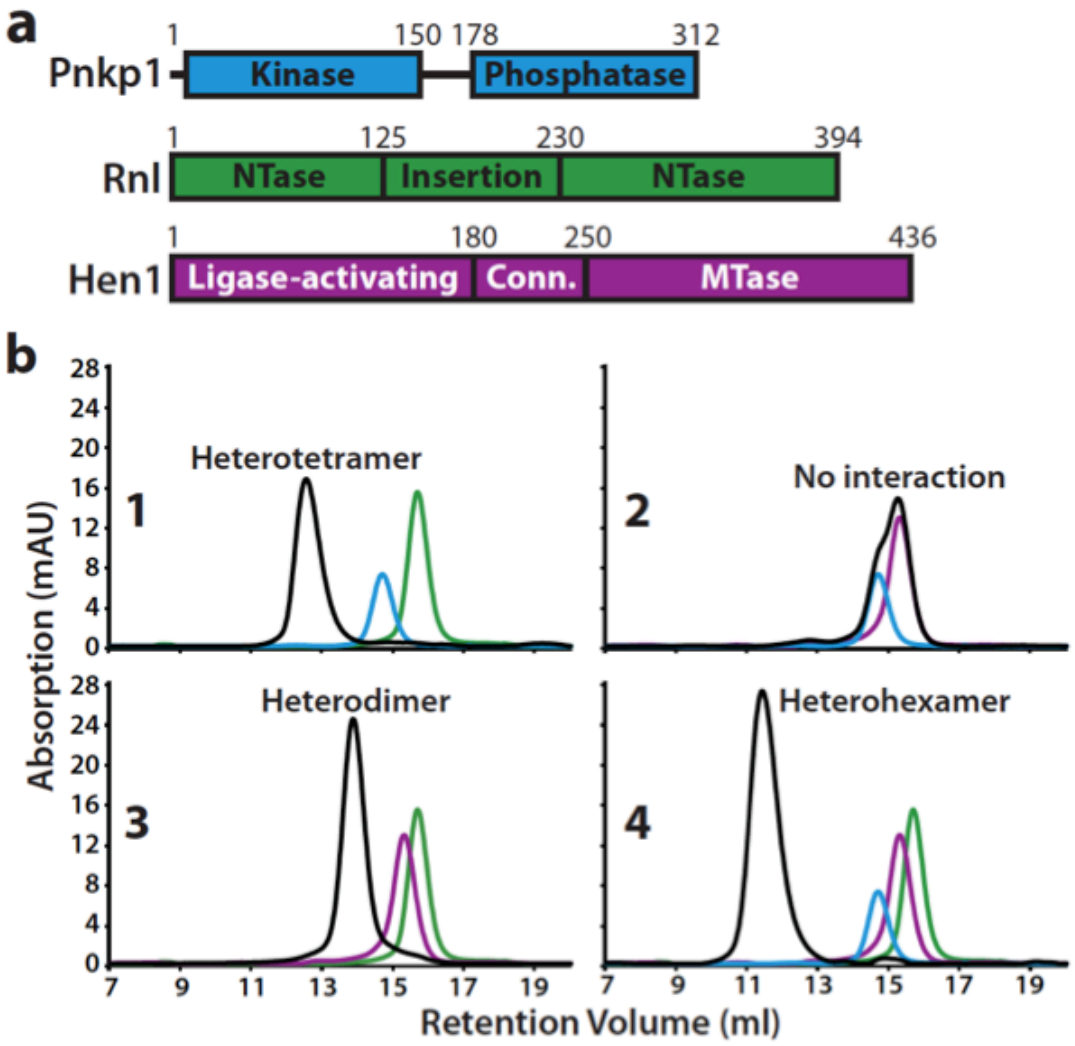

Figure 3.3 Complex formation of *Cgi*Pnkp1/Rnl/Hen. (a) Schematic view of the three proteins that constitute the new bacterial RNA repair system. The boundary of domains in each protein was determined based on the structure of the Pnkp1/Rnl/Hen1 heterohexamer. NTase, nucleotidyltransferase domain; Conn, connecting domain; MTase, methyltransferase. (b) Size exclusion chromatography analyses of individual protein, the pair-wise mixture, and the three-protein mixture. The chromatographic curves of individual protein are colored the same as in a, and the ones for the mixtures are in black.

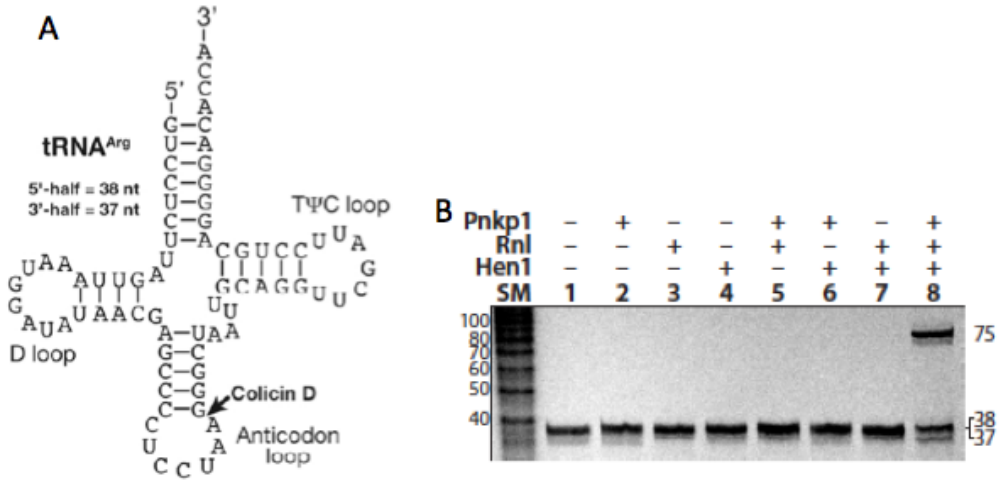

Figure 3.4 Repair of the colicin D-cleaved tRNA^{Arg} from E. coli. (A) Secondary structure of the tRNA^{Arg}. The colicin D cleavage site is marked with an arrow. (B) DPAGE analyses of RNA repair of the cleaved tRNA^{Arg}.

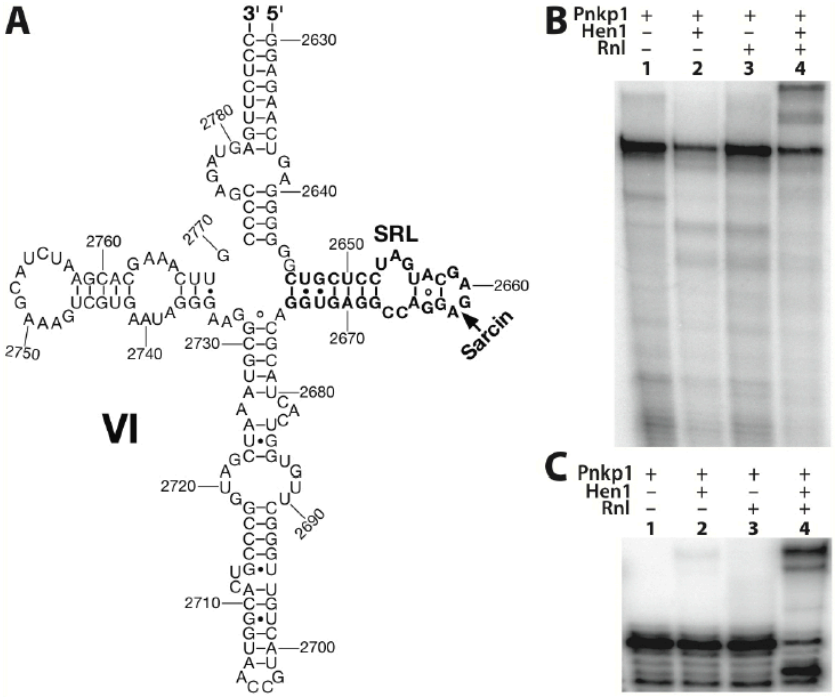

Figure 3.5 Repair of the sarcin-cleaved VI region of 23S rRNA from E. coli. (A) Secondary structure of the VI region of E. coli 23S rRNA. The universally conserved sarcin-ricin loop (SRL) is highlighted in bold, and the sarcin cleavage site is marked with an arrow. (B, C) DPAGE analyses of RNA repair of the cleaved 23S rRNA-VI (B) and SRL (C) by the *Cgi*Pnkp1/Hen1/Rnl RNA repair complex.

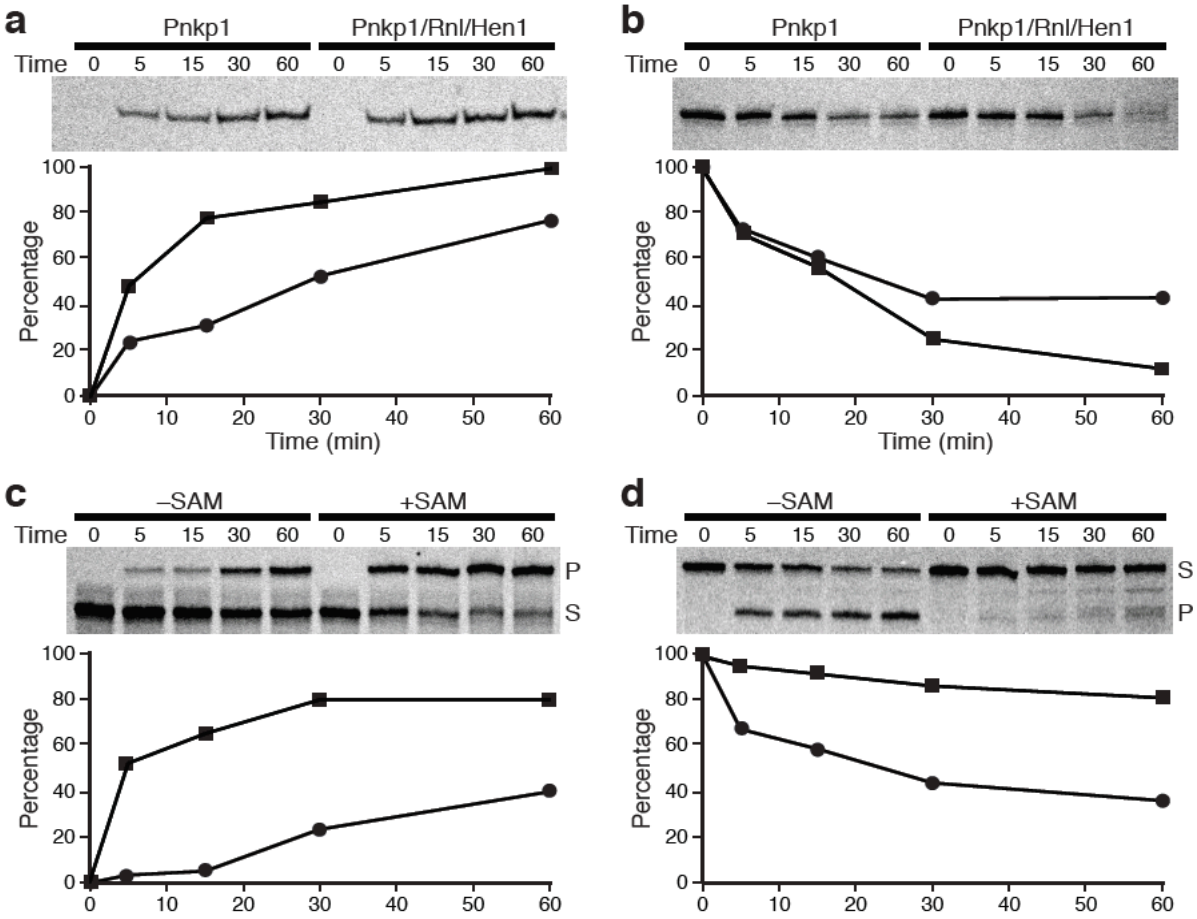

Figure 3.6 Time-dependent assays of enzymatic activities of kinase (a), phosphatase (b), ligase (c) of *Cgi* RNA repair proteins, and nuclease of colicin D (d). The curves marked with cycle represent quantitation of the left panel, and the ones marked with square denote right panel. The ratios of substrate to enzyme are 400 for 5'-phosphorylation reaction shown in a, 4 for 3'-dephosphoylation reaction shown in b, 4 for ligation reaction shown in c, and 1 for cleavage reaction shown in d. S, substrate; P, product. The curves in (a) count the percentage of phosphoryaltion, and the last lane was set as 100%; and the curves in (b) count the percentage of RNA with phosphate at 3'-end, and the first lane was set as 100%; curves in (c) count the percentage of ligated RNA; and curves in (d) count the percentage of uncleaved RNA.

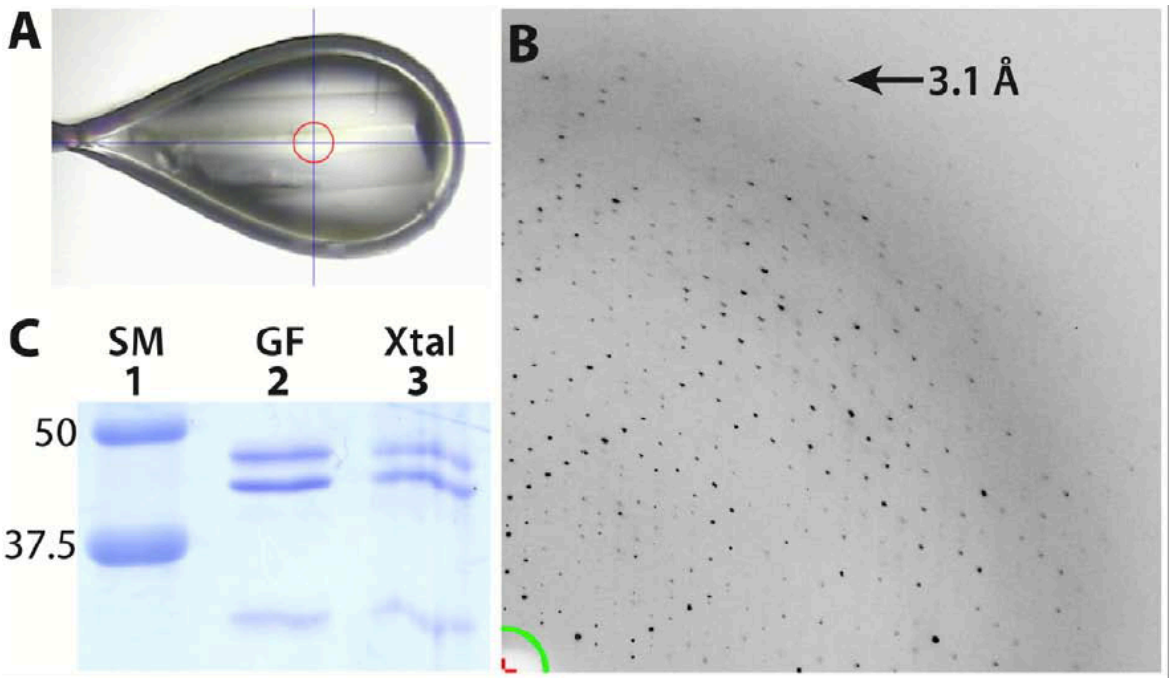

Figure 3.7 Crystallization of *Cgi*Pnkp1/Hen1/Rnl heterohexamer. (A) An image of a crystal in a nylon loops. The red circle denotes the X-ray beam (50 μm). (B) A quarter of a X-ray diffraction map from the crystal shown in (A). Observed diffraction at 3.1 Å is marked with an arrow. (C) SDS gel analysis of the six dissolved crystals (Xtal) along with the gel filtration sample (GF).

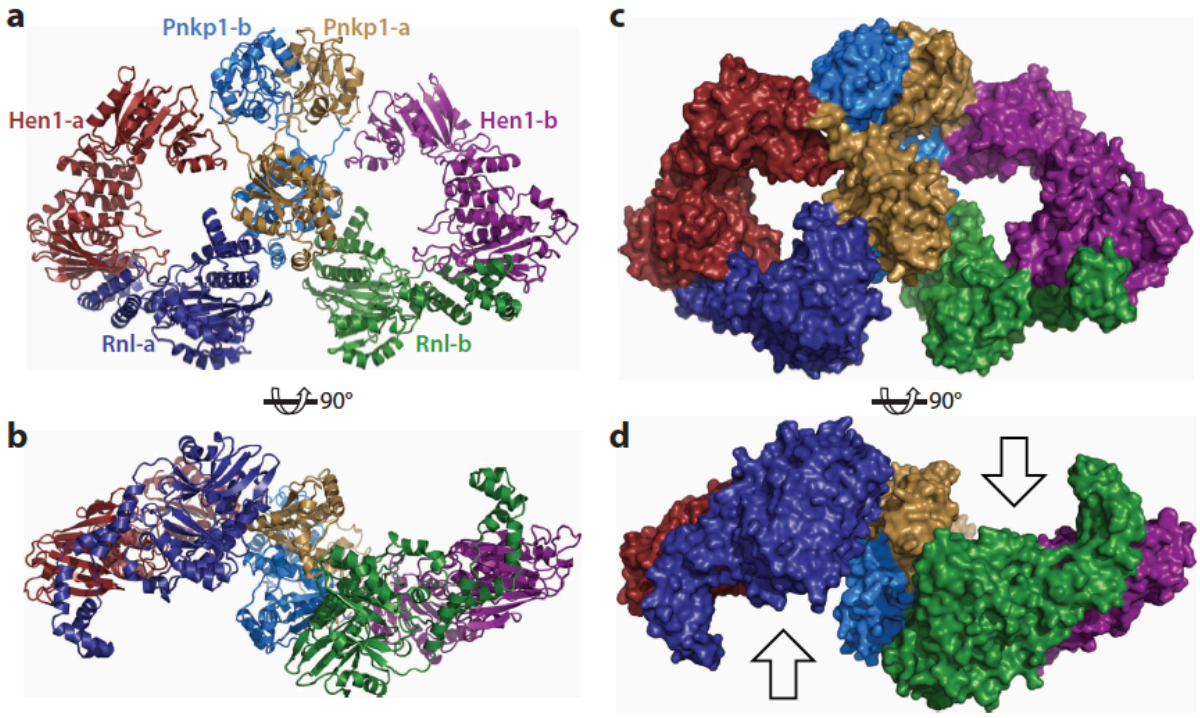

Figure 3.8 Overall structure of Pnkp1/Rnl/Hen1 heterohexamer. (a,b) Ribbon representation of the top (a) and the side (b) views of the structure. One copy of Pnkp1, Rnl, Hen1 are colored the same as in Figure 3.3a, and the second copy of Pnkp1, Rnl, Hen1 are colored sand, darkblue, ruby, respectively. (c,d) Surface of the Pnkp1/Rnl/Hen1 heterohexamer in the same colors and orientations as in a and b, respectively. Two arrows in d indicate the likely directions from which the damaged RNAs approach the Pnkp1/Rnl/Hen1 heterohexamer for RNA repair.

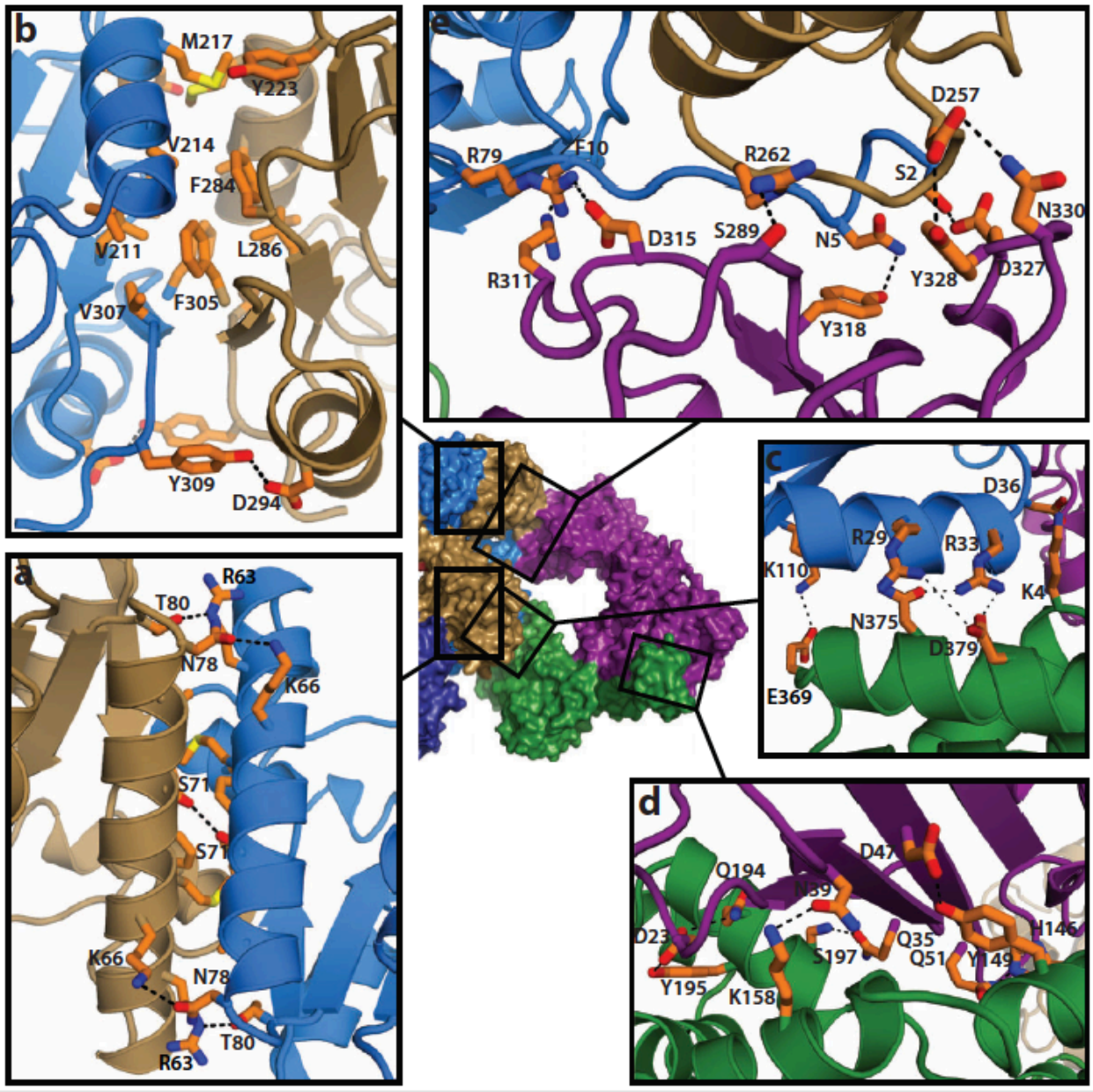

Figure 3.9 Details at the protein-protein interfaces. (a) Details of the interface between the two kinase domains of the Pnkp1 homodimer. Cα-chains of the structure are represented and colored the same as in Figure 3.8a. The side chains are in stick and colored orange except heteroatoms, which are colored individually (nitrogen in blue, oxygen in red, and sulfur in yellow). Hydrogen bonds and salt bridges are depicted with black dashed lines. (b-e) Details of the interface between the two phosphatase domains of the Pnkp1 homodimer (b), Pnkp1 and Rnl (c), Rnl and Hen1 (d), and Pnkp1 and Hen1 (e).

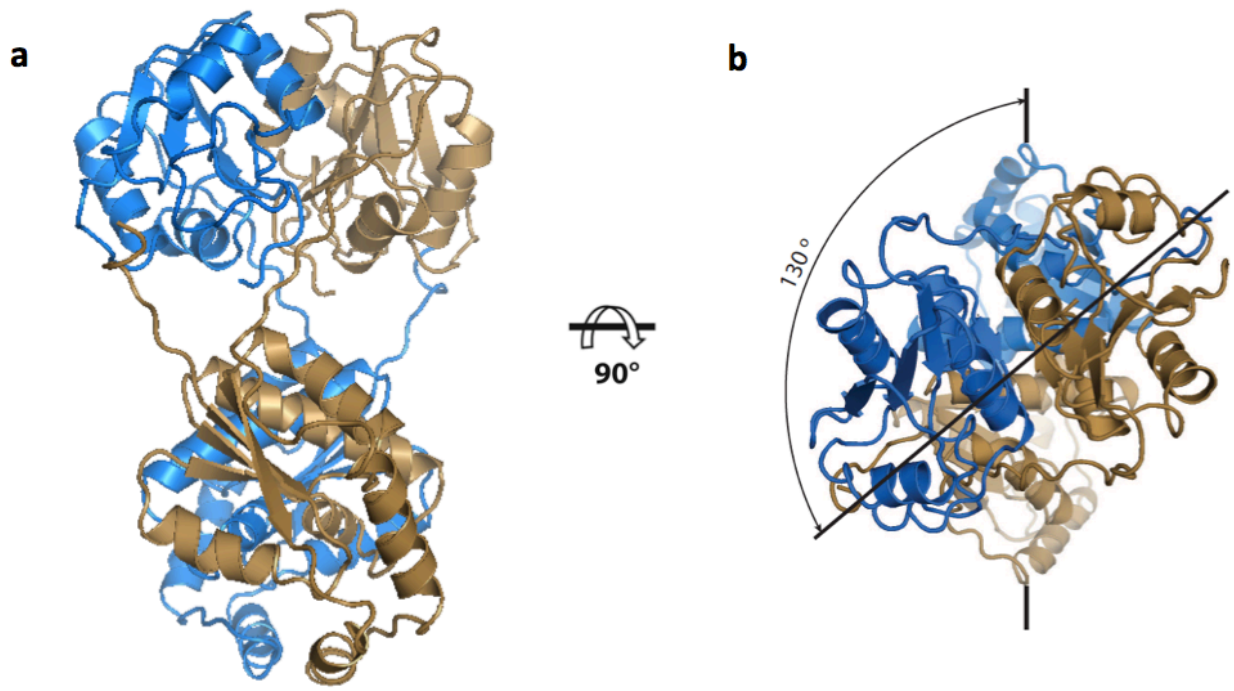

Figure 3.10 Structure of *Cgi*Pnkp1 homodimer in the RNA repair complex. (a) Pnkp1 homodimer in the structure of Pnkp1/Rnl/Hen1 heterohexamer (Figure 3.7a) is highlighted (Pnkp1-a is colored sand, and Pnkp1-b is colored cyan), and the dimmerization of Pnkp1 mainly occurs between the same type of domains (two phosphatase domains at the top; and two kinase domains at the bottom) from both copies. (b) Top view of Pnkp1 homodimer reveals the relative orientations of the kinase and phosphatase modules of the Pnkp1 homodimer observed in the structure of the Pnkp1/Rnl/Hen1 heterohexamer.

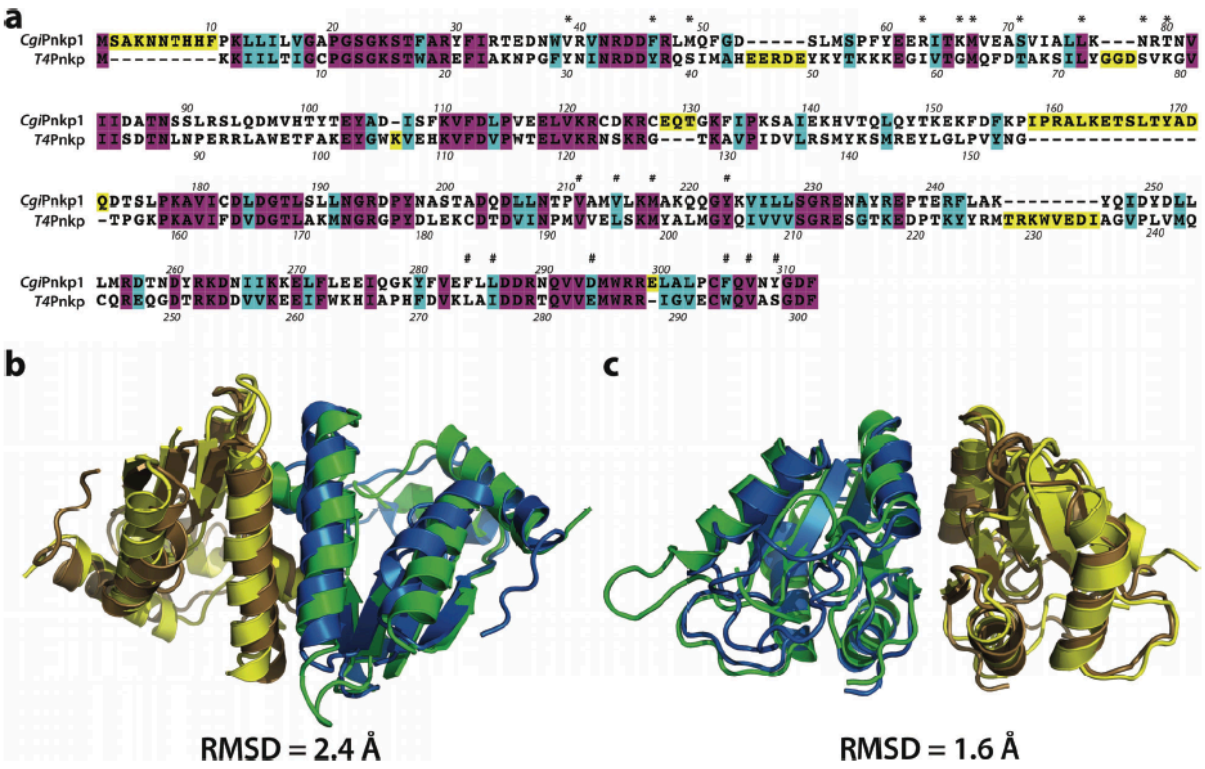

Figure 3.11 Sequence and structural comparison of *Cgi*Pnkp1 with *T4*Pnkp. (a) Amino acid sequences of *Cgi*Pnkp1 and *T4*Pnkp were aligned. The conserved residues are boxed in color, with completely conserved residues in magenta and similar residues in cyan. Residue numbers above the alignment belong to *Cgi*Pnkp1, and those below correspond to T4Pnkp. The residues marked with asterisks and number signs are involved in dimerization of kinase and phosphatase domains of *Cgi*Pnkp1, respectively. (b) Structural alignment of the kinase module of *Cgi*Pnkp1 (residues 2-157) with the corresponding domains in *T4*Pnkp (Accession code: 2IA5, residues 2-154). *Cgi*Pnkp1 are colored and oriented the same in Figure 3.8a, and T4Pnkp are colored yellow and green, respectively. (c) Structural alignment of the phosphatase domains of *Cgi*Pnkp (residues 173-312) with the corresponding domains in *T4*Pnkp (Accession code: 2IA5, residues 155-301). The structures are colored the same as in b and oriented the same as in Figure 3.8b.

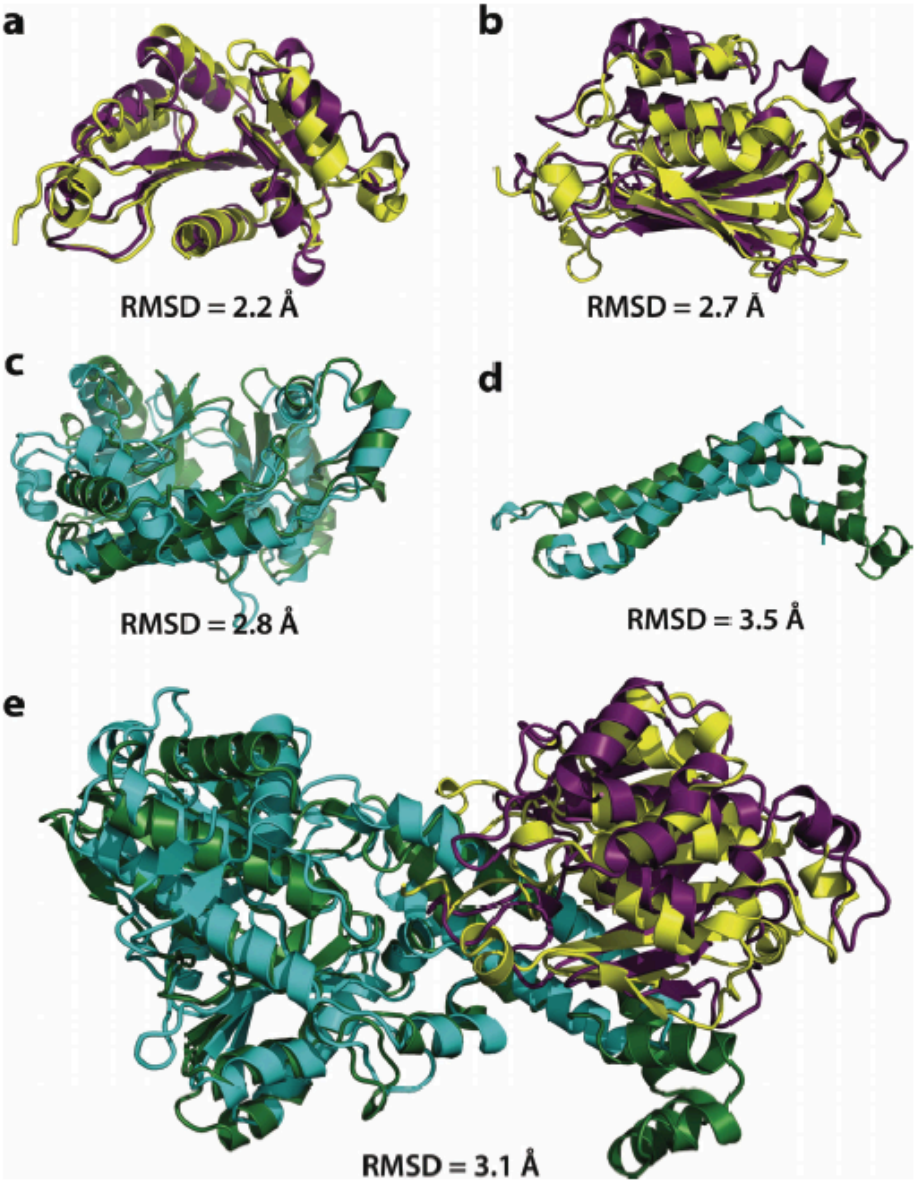

Figure 3.12 Similarity and difference of the structures between Pnkp1/Rnl/Hen1 and Pnkp/Hen1 RNA repair complexes. Structural alignments of the C-terminal methyltransferase domains of Hen1 (a), the N-terminal ligase-activating domains of Hen1 (b), ligase domains of Rnl and Pnkp (minus the insertion domains) (c), the insertion domains (d), and the entire ligase modules of Rnl/Hen1-N and Pnkp- C/Hen1-N (e). The structural components of the Pnkp1/Rnl/Hen1 system are depicted and colored the same as in Figure 3.8, and those from the Pnkp/Hen1 system are colored green (Pnkp) and yellow (Hen1), respectively. The comparisons are based on the structure of the C-terminal methyltransferase domain of Hen1 from *Clostridium thermocellum* (*Cth*) (Accession code: 3JWI), and the structure of the ligase module of *Cth*Pnkp/Hen1 (Accession code: 4E6N).

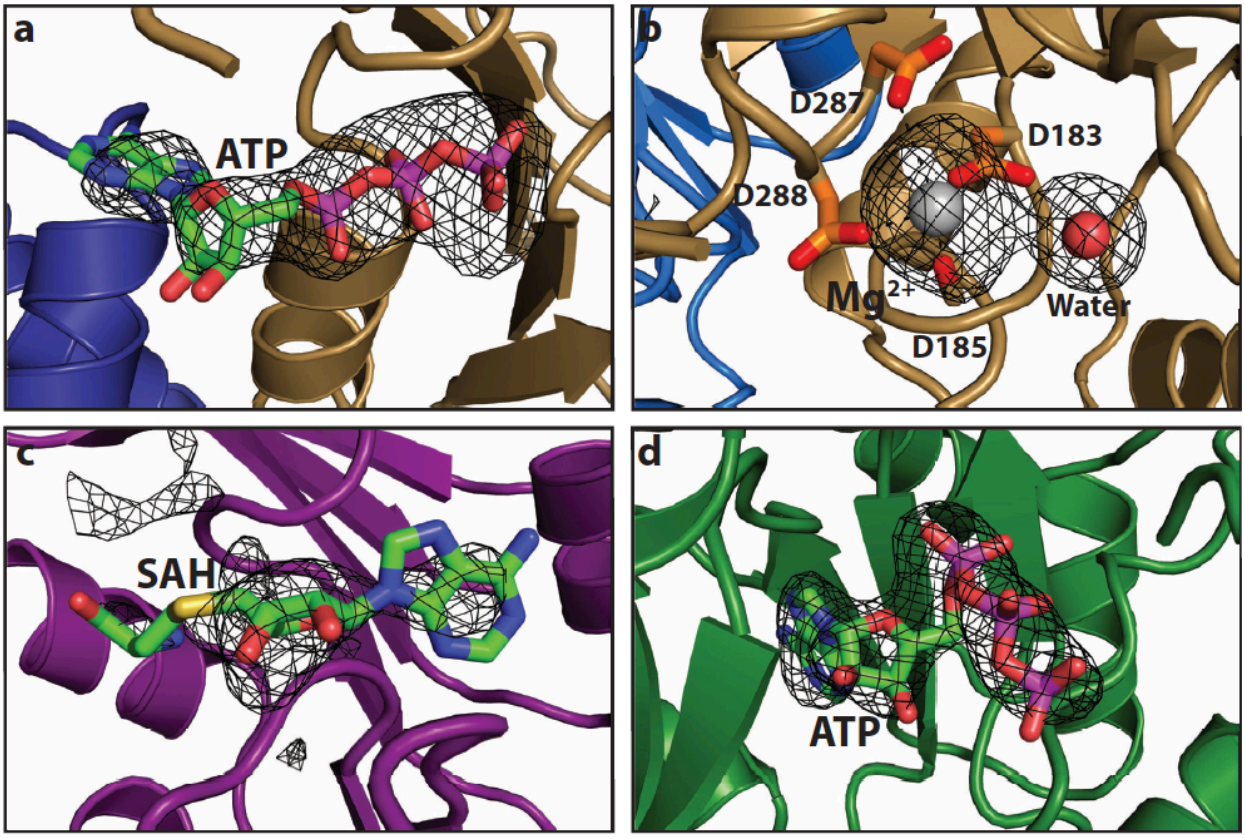

Figure 3.13 Presence of cofactors in all four different active sites required for RNA repair. (a) ATP bound in the kinase active site of Pnkp1-a. (b) A magnesium ion bound in the phosphatase active site of Pnkp1-a. A water molecule was tentatively modeled on additional electron density 4.5Å away from Mg^{2+}, which could also be a phosphate group with a reduced occupancy. (c) SAH was modeled in the methyltransferase active site of Hen1-b. (d) ATP bound in the ligase active site of Rnl-b. Proteins are depicted and colored the same as in Figure 3.8a. ATP and SAH are in stick and colored green with the exception of heteroatoms, which are colored individually (nitrogen in blue, oxygen in red, phosphate in magenta, and sulfur in yellow). Mg^{2+} and H_2O are depicted in sphere and colored silver and red, respectively. All cofactors are covered with simulated annealing composite 2mFo-DFc omit density maps contoured at 1.5σ. Because SAM or SAH was not added during crystallization or crystal soaking, the SAH modeled in the methyltransferase active site was likely obtained by Hen1 during its overexpression in *E. coli*. A significantly weaker omit map compared to others indicates that only a small fraction of SAH is retained by Hen1 after protein purification and crystallization steps.

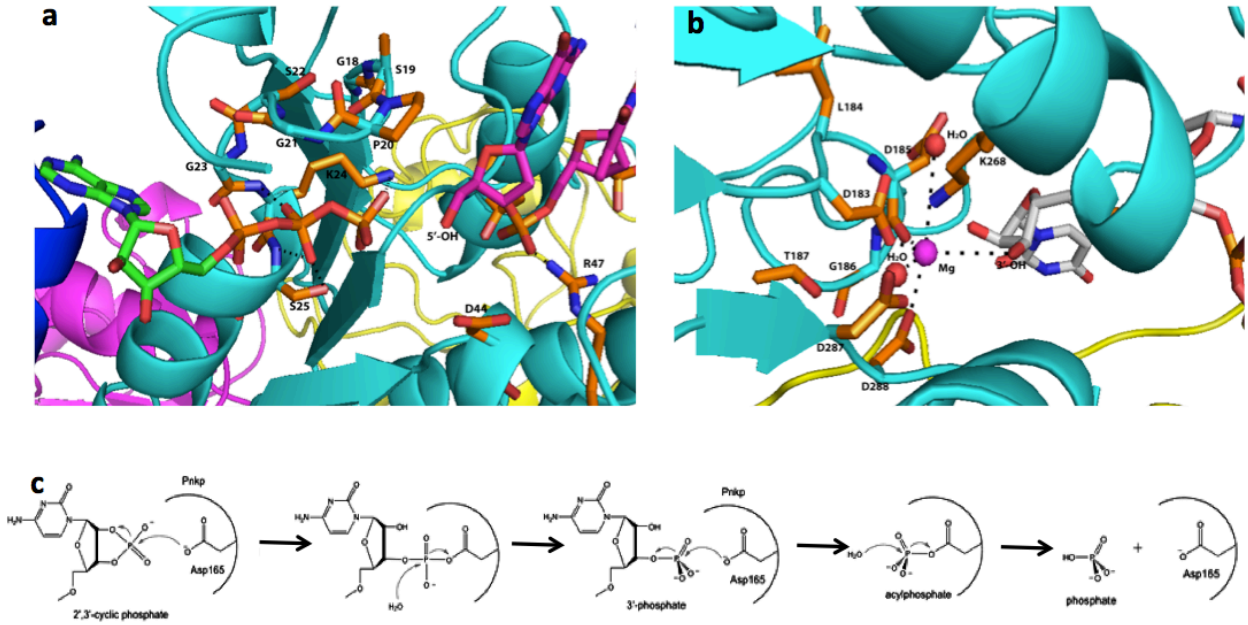

Figure 3.14 The proposed mechanism of kinase and phosphatase of *Cgi*Pnkp1. (a) The P-loop (GXXXXGK(T/S)) of kinase domain (colored in cyan) is shown in stick (colored in the same way as Figure 3.8), and supposed to bind ATP (carbon is colored green). A polynucleotide substrate (carbon is colored magenta) is docked in this active site, It is shown that Arg47 can help stabilize the 3'phosphate of the substrate while Asp44 activate 5'-OH of nucleonic acid by abstracting its proton and facilitate the attack to gamma phosphate of ATP. (b) The DxDxT motif of phosphatase domain (colored in cyan) is shown in stick. The metal ion is shown as sphere (magenta) in active site, and it can be coordinate with side chain and main chain of the motif, two water molecules, 2', 3'-PO_4 of docked polynucleotide substrate (carbon is colored white). (c) The mechanism of phosphatase in *T4*Pnkp (75). Phosphatase converts a 2', 3'-cyclic phosphate to 3'-OH, 2'-OH ends through sequential diesterase and monoesterase reaction. Asp183 in *Cgi*Pnkp1 is required for the phosphate removal through covalent asptyl-phosphate intermediate, while Asp185 donates a proton to the leaving group, ribose O3.

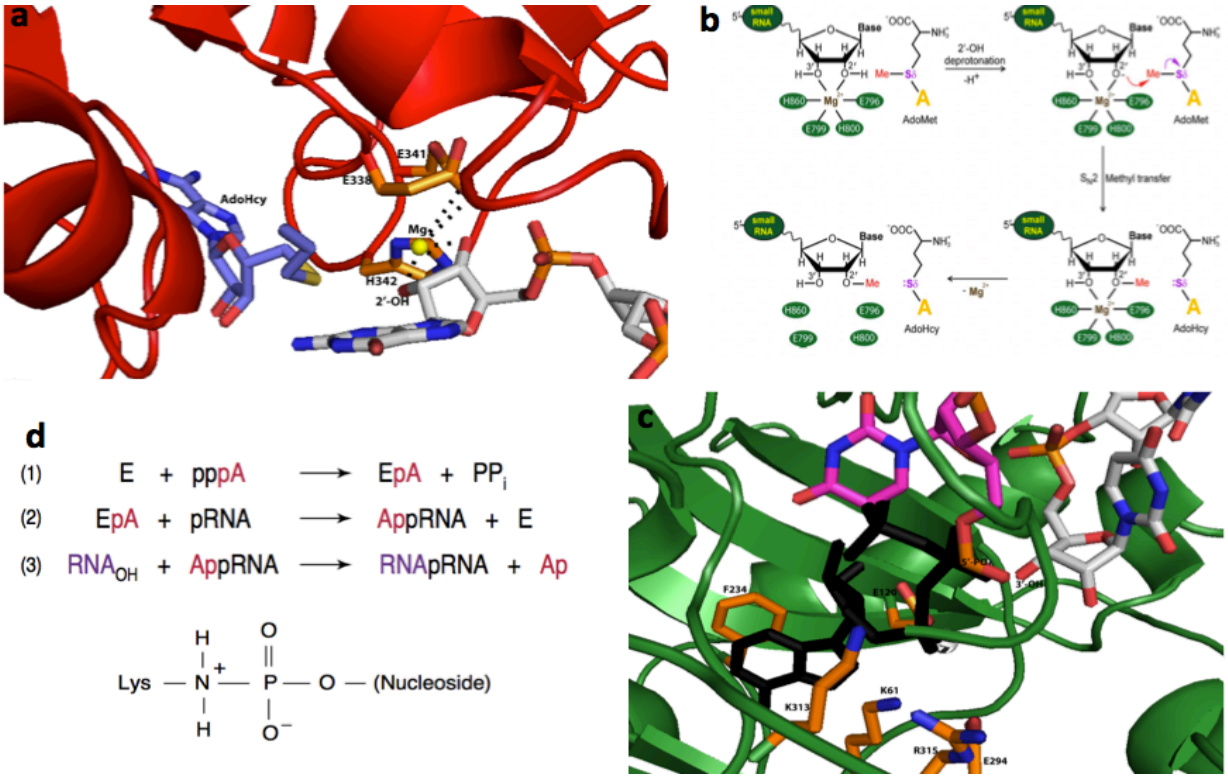

Figure 3.15 The proposed mechanism of methyltransferase and ligase. (a) The active site of methyltransferase (red) shows the density of SAH (carbon is colored purple), the demethylated product of SAM. The docked Mg^{2+}(yellow sphere) could be coordinated by the 2'-OH, 3'-OH of RNA substrate (carbon is colored white), as well as the side chain of conserved residues (E338, E341, H342, shown in stick and colored in the same way as Figure 3.8); (b) The proposed mechanism for 2'-O methyltransferase at the 3'-end of RNA substrate (77), in which the conserved critical residues are highlighted in green and the Mg is required for the reaction and stabilized only in the presence of RNA substrate; (c) Two strands of RNA (the carbon of 5'-half is colored magenta, and the carbon of 3'-half is colored white) are docked in to the active site of ligase domain (colored green). The residues important for substrate binding or catalysis are shown in stick (color is the same as Figure 3.8). ATP was colored black, and its α-phosphate could be attacked by K61 to form the covalent adduct (d, bottom) (59). The remaing steps of ligation (d, top) require the involvement of both strands of RNA.

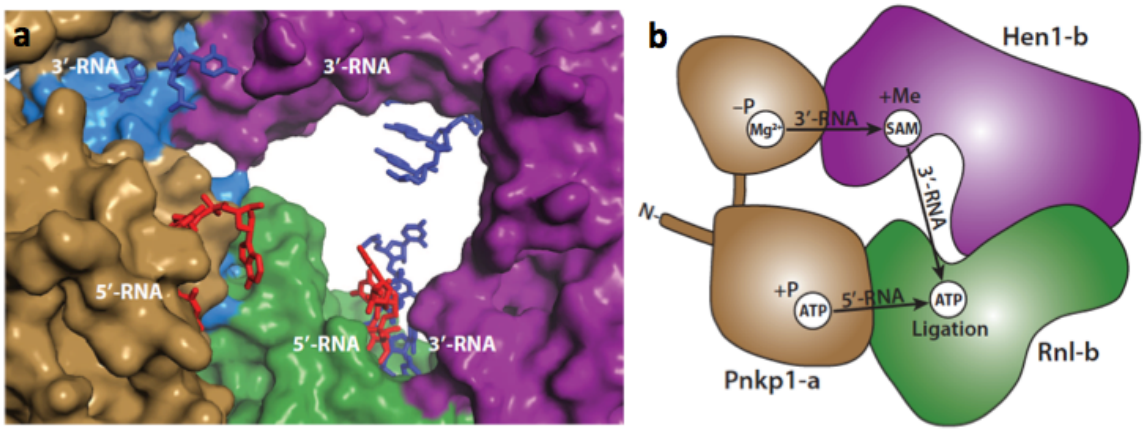

Figure 3.16. Proposed RNA repair pathway for *Cgi*Pnkp1/Rnl/Hen1. (a) Modeling the 5'-end (red) and the 3'-end (blue) of damaged RNAs into all four active sites required for RNA repair. The modeled RNAs are in stick, and the proteins are depicted the same as in Fig. 2c but with a closer view on the vacant space surrounded by the b-unit of the Pnkp1/Rnl/Hen1 heterotrimer (colored cyan, green and magenta) and Pnkp1-a (colored sand). (b) A proposed mechanism of RNA repair carried out by the Pnkp1/Rnl/Hen1 heterohexamer. Pnkp1-b is omitted for clarity. The locations of the active sites are indicated by cofactors (white cycles), with ATP in the kinase and ligase active sites, a magnesium ion in the phosphatase site, and SAM in the methyltransferase site. Arrows indicate the travel pathways for both 5'-end and 3'-end of a damaged RNA for RNA repair.

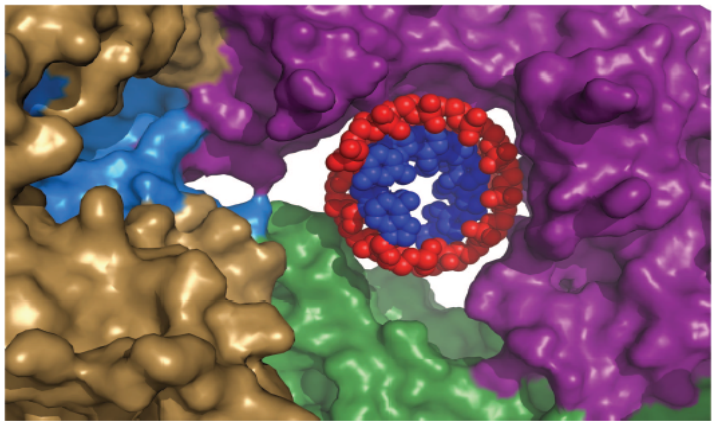

Figure 3.17. Measuring the size of the vacant space at the center of the ring structure with a double-stranded RNA (dsRNA). The standard dsRNA was created in Coot program and was manually docked into the vacant space surrounded by the b subunit of the Pnkp1/Rnl/Hen1 heterotrimer and Pnkp1-a (color is the same as Figure 3.15).

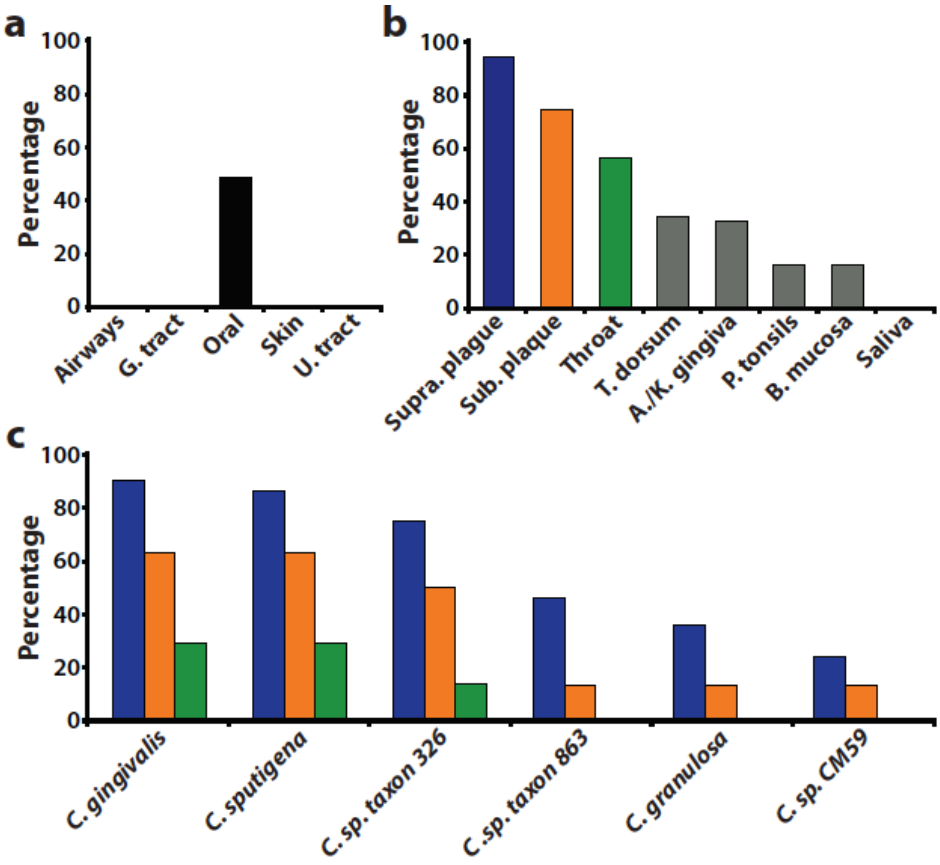

Figure 3.18. Presence of the Pnkp1/Rnl/Hen1 RNA repair complex in human-host microbes. (a) Distribution of bacteria possessing the Pnkp1/Rnl/Hen1 RNA repair complex at five different locations of human body. (b) Distribution of the bacteria at different sub-locations of human mouth. (c) Distribution of six human-hosted bacterial species possessing the Pnkp1/Rnl/Hen1 RNA repair complex at three major sub-locations of human mouth.

Table 3.1 Data collection and refinement statistics for the crystals in chapter 3

	Pnkp1/Rnl/Hen1 Hexamer + Cofactors	Pnkp/Rnl/Hen1 Hexamer
Data Collection		
Data type	SAD	Native
Space Group	$P2_1$	$P2_1$
Cell dimensions		
a, b, c (Å)	108.5, 187.2, 112.0	111.2, 179.2, 114.3
α, β, γ (°)	90.0, 106.0, 90.0	90.0, 103.7, 90.0
Resolution (Å)	50.0-3.4	50.0-3.3
Wavelength (Å)	0.9794	0.9998
Redundancy	7.5	5.1
Completeness (%)	98.4 (91.1)	99.8 (98.7)
$I/\sigma I$	17.1 (3.1)	13.9 (3.8)
R_{merge} (%)	4.4	7.0
Refinement		
Resolution (Å)	42.9-3.4	47.2-3.3
No. reflections	57,526 (5,292)	65,456 (6,432)
R_{work}/R_{free} (%)	17.6/23.3 (24.8/31.4)	17.3/23.7 (25.1/30.7)
No. atoms	18,713	18,886
Protein	18,442	18,711
Ligand/ion	179	84
Water	92	91
B factors (Å^2)		
Protein	85.4	71.3
Ligand/ion	109.1	97.6
Water	54.9	48.7
r.m.s. deviations		
Bond lengths (Å)	0.010	0.012
Bond angles (°)	1.55	1.64

*Values in parenthesis are for highest resolution shell.

CHAPTER 4: IN VIVO STUDIES OF RNA DAMAGE AND REPAIR

1. Introduction

Although we have characterized the biochemical function and structural features of RNA repair complex, we still have little knowledge of the biological function for RNA repair. The bacteria are exposed to various toxins in the wilderness, which is difficult to replicate in a laboratory setting. Since we have previously worked on the ribotoxins that target essential RNA such as tRNA and rRNA and provide potential substrates for RNA repair system, we decided to introduce similar toxins as well as RNA repair system into bacteria to evaluate if RNA damage and repair could be observed *in vivo*.

1.1 Ribotoxins

Comparative genomic analysis classifies the prokaryotic defense systems into two major groups (Figure 4.1) (82, 83): 1) The first group, including R-M (resitriction-modification) (84-86), DND (DNA phosphorothioation) (87) and CRISPR (clustered regularly interspaced short palindromic repeats) (88,89), works on non-self invader; 2) the second group, including TA (toxin-antitoxin) (90,91), ABI (abortive infection) (92), functions on the cell death or dormancy The majority of toxins in the second group interferes with the translation process mostly through RNA cleavage, and is classified as ribotoxin.

That toxin molecules are produced and deployed by organisms is the most common strategy in inter-organismal or intra-genomic conflicts. The toxins are highly diversified, spanning an entire biochemical spectrum from diffusible small molecules to proteinaceous toxins (93). The proteinaceous toxins are important in every level of biological conflict, such as the conflict between different eukaryotes (e.g. sarcin and racin), multicellular organisms against their

pathogens (e.g. RNaseA and RNaseL) and inter-specific conflict in bacteria (e.g. secreted toxins, such as colicin D).

Toxins employ a variety of mechanisms to cause cellular damages, but the effects can be classified into two major categories: One is the disruption of cellular integrity (e.g. forming pores in membrane) (94). The other is the use of enzymatic activities to block the flow of biological information by targeting nucleic acids or proteins (e.g. targeting translation apparatus) (11-23). In this study, we focus on the toxins belonging to the second type, which represents the most prominent theme of cellular toxicity and is more closely related to the antidotes such as RNA repair. The known and predicted ribotoxins used for the following assays mostly target essential RNAs involved in protein translation.

Many kinds of proteinaceous toxins target ribosomal RNAs, which usually results in cell death through inhibition of protein synthesis. For example, ricin and sarcin target the Sacin-Racin loop (SRL) of eukaryotic ribosome, which prevents EFTu from binding (11-13). In prokaryotes, VapC20 from *Mycobacterium tuberculosis* was recently discovered to also cleave SRL at the same position as sarcin (14). Unlike sarcin that generates 2', 3'-cyclic phosphate group, however, VapC20 belongs to PIN (pilT amino-terminal) family of protein (95), which generates 3'-hydroxyl group.

In addition to rRNA, transfer RNAs constitute another universal and essential RNA component required for protein synthesis. Relevant to this study here are Colicin D produced by *E. coli* cells (22) and VapC from *M. tuberculosis* (23). They both contain genes encoding the proteinaceous toxins that inhibit cell growth as well as the antitoxins that counteract the toxins via direct protein-protein interaction. Each toxin has its specificity to recognize distinct features of specific tRNA and cause the cleavage at the anticodon loop, resulting in inhibition of global

translation. Colicin D cuts isoacceptors of $tRNA^{Arg}$ between position 38 and 39, while VapC recognizes and cleaves initiator tRNA at the wobble position in a divalent metal ion-dependent manner and generates the same ends as VapC20 does.

1.2 *In vivo* toxicity assays

To test the effect of proteinaceus toxins on bacteria, we chose the tightly controlled plasmid, which expresses little protein in the absence of an inducer. A series of plasmid vectors have been constructed, which contain the P_{BAD} promoter of the *araBAD* operon and the gene encoding the positive and negative regulator of this promoter, *araC*. In the absence of arabinose, *araC* gene product AraC binds to and loops the regions upstream of the P_{BAD} promoter, preventing the binding of RNA polymerase. However, AraC would change its conformation upon the arabinose binding and behave like an activator by loosing the loop structure and facilitating the binding of RNA polymerase to the promoter (Figure 4.2). The ratio of induction/repression allows the regulation in the 1,200- fold range, and the responses of induction and repression are very rapid (96).

1.3 *In vivo* RNA repair

In order to test RNA repair *in vivo*, we selected another vector expressing RNA repair proteins. The second vector is also required to be tightly controlled, but compatible with the one containing toxin gene. In addition, compared to the first vector, the second vector should use different inducer and exhibit different antibiotic resistance. We chose the PRO bacterial expression system for the expression of RNA repair proteins, for this system meets all the

requirement and exhibits two-step induction with the addition of IPTG and arabinose (Figure 4.2) (97).

2. Methods and Materials

2.1 Plasmids and bacterial strains

The pBAD plasmids with chloramphenial resistance were used to express the toxin genes, whereas PRO bacterial expression system with kanamycin resistance were used for RNA repair proteins expression (all the plasmids were provided by Xinyun Cao from Dr. John Cronan's lab). The *E.coli* strain used in this study is DH5α.

2.2 Test of bacterial growth

The cells transformed with toxin genes or/and RNA repair genes were cultured at 37°C overnight. The culture was adjusted by autoclaved water until the OD_{600} is approximately 0.1. The ajusted culture was further diluted 10 and 100 times, respectively, and 2ul of each diluted sample was spotted on the plates with indicated inducer or antibiotics.

2.3 RNA isolation and gel electrophoresis

RNA was extracted from the cells via hot phenol/chloroform method. First, the cells were recovered by centrifuging 3 mL of cell culture to remove the liquid. Then cell pellet was resuspended in 200 μl ice-cold solution I (0.3M sucrose, 0.01M Na-acetate, pH 4.5) and rapidly mixed with 200 μl solution II (2% SDS, 0.01M Na-acetate, pH 4.5) and 400 μl phenol (pH 4.5). Sample was then vortexed for 1 min and incubated at 65°C for 3 min. Then the sample was submerged into liquid nitrogen and centrifuged for 5 min at room temperature. The top phase

(~400ul) was taken out and mixed with 400ul phenol. The steps of vortexing, heating, freezing and centrifugating were repeated as described. The top phase was taken out and mixed with 400 μl chloroform. After vortexing and centrifugation, around 270ul top phase was transferred to the new tube, and the RNA was precipitated via ethnol precipitation. The RNA pellet was washed by ice-cold 70% ethanol and dried at room temperature. The dry RNA was resuspended in 20 μl DPAGE loading buffer, heated at 95°C for 3 min. 7 μl of each RNA sample was loaded into 5% denaturing polyacrylamide gel. The bands corresponding to RNA were visualized by ethidium bromide staining.

2.4 Northern blotting

DNA probe was ordered from DNA IDT, and labeled at 5'-end by [γ^{33}P]-ATP. In a 20 μl reaction volume (25 mM Tris pH 8.0, 50 mM KCl, 25 mM MgCl2, 0.05 mM EDTA, 5 mM DTT, 0.25 mM $MnCl_2$), 42 pmols DNA was incubated with 5 μl [γ^{33}P]-ATP (3000Ci/mmol, 50uCi used, PerkinElmer) and 3 μl *Cth*Pnkp-N (10uM) at 37 °C for 1 hour. 2 μl 0.5M EDTA was added into the reaction to quench the reaction. Micro Bio-Spin 6 column (BIO-RAD) was used to remove the free ATP. The purified DNA probe was aliquot and stored in -20°C for future use.

Total RNA was prepared and separated on denaturing polyacrylamide gel as described in 2.3. The RNA was then transferred to a Hybond-N+ membrane (GE Healthcare) by electroblotting in 1×Trisborate-EDTA buffer (100V, ~1hour). The membrane was cross-linked by UV, and prehybridized in 1× PerfectHyb Plus hybridization buffer (Sigma) for 5min at 42 °C. 1 μl probe was added in the hybridization buffer and hybridization was incubated overnight at 42°C. The membrane was washed three times by 2× saline-sodium-citrate buffer (National

diagnostics) with 0.1% SDS. The radioactivity was detected by a PhosphorImager system (Molecular Dynamics).

3. Results and Discussions

3.1 Discovery of potential new ribotoxins

We employed the follwing three different approaches to discover new potential ribotoxins.

First, we used amino acid sequences of VapC, VapC20 genes to carry out BlastP search and found three ribotoxin candidates (A4YDT6, A4YF78 and D0KNK4) from *Sulfolobus solfataricus*, which is a thermophilic archaeon and has been well studied for the mechanism of DNA replication, RNA processing, transcription, translation and so on. We also found a VapC homolog in *Capnocytophaga gingivalis*, the organism providing genes for our RNA repair study as described previously used for RNA repair study previously (95).

Second, BECR (Barnase-EndoU-Colicin E5/Colicin D-RelE) like nuclease fold was reported to constitute a large assemblage of metal-independent RNase (8, 10, 15). The common structural unit of catalytic domain in these toxins contains a N-terminal helical segment followed by a 4-stranded β sheet. Most of the members contain two conserved histidines and one conserved alcoholic residue S/T for enzymatic activity. We found Colicin D-like domain in *Pseudomonas aeruginosa* and *Salinispora tropica*. Moreover, an RNase candidate of the BECR fold related to Colicin D clade is also found consistently associated with flagellar operon. It was proposed that the RNase might be delivered by the flagellar system or RNase worked as RNA-processing enzyme to regulate flagellar gene expression. We selected one such candidate from

Clostridium thermocellum, which has also been used for providing us RNA repair genes for our previous study.

Third, Tox-EDA39C, named after the plant protein EDA39, was reported as another uncharacterized toxin domain found in polymorphic toxin system of a wide range of bacteria (93). The presence of this domain in several eukaryotic lineages suggests that it might have been acquired by eukaryotes from bacterial endosymbionts and recruited as a potential RNase for anti-pathogen defense. The conservation pattern of catalytic residues for this domain is similar to that in RNase A domain, which indicates a novel metal-independent RNase mechanism. We selected a candidate of EDA39 domain from one of *E.coli* strains.

3.2 The effect of toxins on bacterial growth

3.2.1 The toxicity of known toxins

We first tested the toxicity of known toxins colicin D, sarcin, VapC, VapC20 in order to confirm that the system for our assay works as expected. Thus, the genes encoding these toxins were inserted individually into the plasmid after P_{BAD} promoter, which could be induced by arabinose. The plasmids were transformed into DH5a cells, and the overnight culture of those cells was adjusted according to the cell density before spotted on the plates with or without arabinose. We found that colicin D, VapC and VapC20 were able to efficiently kill the *E. coli* cells but sarcin was less effective (Figure 4.3). Sarcin was found to cleave the ribosomal RNA in eukaryotic cells efficiently. Therefore, our data indicates that it might be less effective with the bacterial ribosomal RNA.

3.2.2 Confirmation of new ribotoxins

As described in section 3.1, we selected eight ribotoxin candidates, three from Sulfolobus solfataricus (A4YDT6, A4YF78, D0KNK4), two from the species containing RNA repair system (*Cgi*-Pin, *Cth*-BECR), the Colicin D-like domain from the *Pseudomonas aeruginosa* and *Salinispora tropica* (*Pae*-CD, *Str*-CD), and one uncharacterized toxins domain from *E.coli* (*Ec*-EDA) for our assays. These genes were inserted into the plasmid after P_{BAD} promoter and their toxic effects were assayed analogous to the ones of known ribotoxins described in the previous section. Among all these candidates, three of them (*Cgi*-Pin, *Pae*-CD and *Ec*-EDA) had lethal effect on *E. coli* cells, killing the cells as efficiently as the known toxins Colicin D, VapC (Figure 4.4). For the remaining five, it is possible that some of them are also ribotoxin, but the protein might not be expressed or the RNA damage was not severe enough to cause the cell death. It is also possible that some of them are not ribotoxins. Further studies are required to assess these possibilities.

3.3 Identification of RNA damage by ribotoxins

3.3.1 Analysis of RNA cleavage by gel electrophoresis

To confirm RNA damage caused by the known toxins (Figure 4.5), we extracted total RNAs from cells after induction of toxin genes. The total RNA was analyzed by 5% denaturing polyacrylamide gel. Cells with sarcin and VapC20 treatment resulted in the same length of fragment (~243nt), confirming that both cleave 23S rRNA at the junction of G2661/A2662 as previously reported. A smaller amount of the fragment could also be observed among the total RNAs extracted from the cells containing VapC in the absencse of inducer. The experiment

revealed that cleavage by sarcin was less effective, consistent with the result of cell-based assays described previously.

We also have tested RNA damage caused by unknown toxins (Figure 4.6). Based on cell-based assays, *Cgi*-Pin, *Ec*-EDA and *Pae*-CD could kill the bacterial cells efficiently. Therefore, we analyzed the total RNA from the cells treated with these three toxins using VapC as positive control (cleave initiator tRNA). First of all, we did not detect fragment of ~243 nt as before, indicating that SRL loop was unlikely the target of those toxins. Second, the induction of *Ec*-EDA produces some RNA fragments shorter than tRNA, indicating *Ec*-EDA might target tRNA for cell killing. Third, we failed to detect any obvious RNA cleavage with the treatment of *Cgi*-Pin and *Pae*-CD. It is possible our approach is not sensitive enough to detect RNA cleavage by these two toxins. Alternatively, *Cgi*-Pin and *Pae*-CD might target components other than RNAs for cell killing. Future studies are required to addresss these unresolved issues.

3.3.2 Identification of RNA cleavage by Northern blotting

It has been shown that SRL is close to the 3'-end of 23S rRNA. Therefore, cleavage of SRL by sarcin and VapC20 at G2661/A2662 junction produce ~250 nt RNA fragment. To detect RNA cleavage more sensitively, we employed Northern blotting. The DNA probe was chosen to base pari with residue 2722–2742 of 23S rRNA and the cleavage products were verified the (Figures 4.7, 4.8) (14). As observed previously, VapC20 shows much higher efficiency of SRL cleavage than sarcin. *In vitro* ribosome cleavage was carried out and the result was visualized by Northern blotting (data not shown). The cleavage of sarcin in reaction solution is more efficient than that in bacterial cell, which might be due to the higher concentration of enzyme and more suitable buffer condition in the laboratory setting. Moreover, we proposed that due to the lack of

cellular response to the toxins in vitro, the cleaved ribosome couldn't be removed, resulting in the high percentage of cleavage.

In the future experiments, we plan to expand the library DNAs probes to include different tRNAs as well as 16S RNA to analyze cleavage products of colincin D, VapC as well as our newly discovered toxins.

3.4 Test of possibility of rescuing bacteria from ribotoxin via RNA repair

3.4.1 The effect of RNA repair on bacterial growth

Our previous study of *in vitro* reconstitution of RNA repair shows that both Pnkp/Hen1 and Pnkp1/Rnl/Hen1 RNA repair system work efficiently on the cleaved tRNA. Here we introduce the genes encoding both toxin and RNA repair proteins into the bacteria to evaluate whether RNA repair can resue cells killed by ribotosins.

Colicin D can kill the cells by cleaving the $tRNA^{Arg}$ isoacceptors, and its expression in pBAD plasmid is under the tight control of the inducer arabinose (Figure 4.3). Pnkp/Hen1 operons from different species were inserted into the pPROlar plasmid, and their expressions were controlled by IPTG and arabinose (Figure 4.2C, 4.2D). The results confirm that cells are killed with the expression of colicin D (Figure 4.9, compare sector 1 for the two plates). However, expression of RNA repair gene with the addition of IPTG and arabinose did not result in the recovery of cell growth. (Figure 4.9, compare sector 2, 3, 4 for the two plates). There are several possibilities for this unexpected result: 1) even though the resistance of bacteria to both chloramphenical (pBAD) and kanamycin (pPROlar) indicates that both plasmids are transformed, it is still possible that the bacterial cells gain antibiotic resistance in some other way than from the plasmid, resulting in the absence of RNA repair system; 2) the plasmids, especially

the ones containing RNA repair genes, might undergo some mutagenesis and result in the insufficient expression of functional proteins; 3) arabinose could serve as the inducer for both toxin and repair genes, which might complicate the experiment design; 4) the relative level of ribotoxin to the repair system might be too high, resulting in that RNA cleavage by ribotoxin overwhelm RNA repair by the RNA repair system. Future studies are required to improve the system as well as fine-tune the relative concetration of ribotoxin and RNA repair system.

3.4.2 Possible repair of damaged ribosome by the RNA repair system

We already showed that sarcin could cleave the bacterial ribosome at SRL, but it was not able to cause the cell death under the current laboratory setting. Instead of monitoring the recovery of cell growth, we decided to test whether the α-fragment caused by sarcin would disappear with the treatment of RNA repair proteins. We grew the *E.coli* cells containing pBAD:sarcin plasmid, and harvested the cells 30 minutes after induction of sarcin. The cleaved ribosome was isolated (98,99) and treated with *Cgi*Pnkp1/Rnl/Hen1 complex, followed by extraction of total RNA and analysis by Northern blotting. We tried a low concentration of magnesium ion, which was reported to force the two subunits of ribosome to fall apart (100, 101), but failed to observe the decrease of cleaved α-fragment. We also tried to mimic the stall ribosome by adding initiator tRNA and mRNA into the reaction in the presence of a high concentration of magnesium ion (102, 103). The result showed that the percentage of cleavage was slightly reduced, which might indicate RNA repair (4.10). The experiment needs to be repeated and the experimental condition needs to be optimized in order to produce firmer conclusion.

Since the naked RNA containing cleaved SRL could be efficiently repaired by *Cgi*Pnkp1/Rnl/Hen1 complex (Figure 3.5), we proposed that the cleaved RNA within ribosome might not be as accessible to the enzyme as naked RNA. Moreover, the experimental condition for naked RNA repair might not be applied to isolated ribosome.

4. Figures and Tables

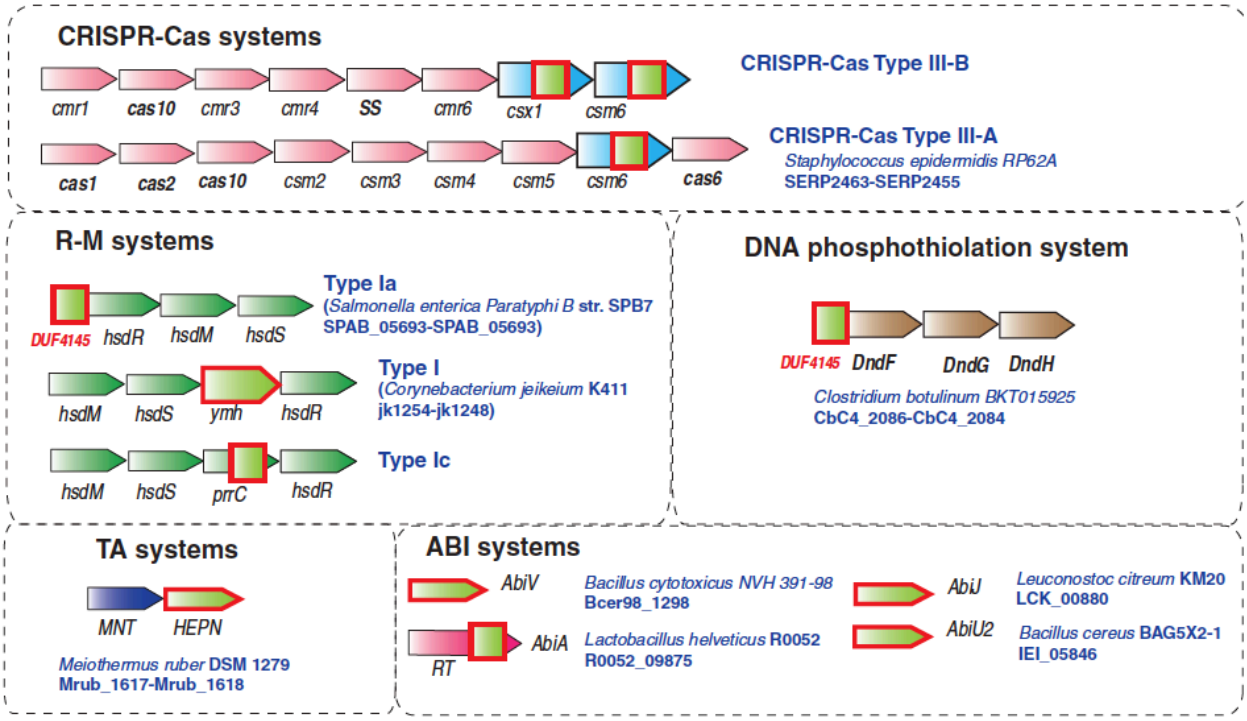

Figure 4.1 Schema of bacterial denfence systems. This figure shows the genomic loci encoding different immunity systems containing HEPN domains (higher eukaryotes and prokaryotes nucleotide-binding domain, predicted endoribonuclease). HEPN domain is indicated by a red outline. (55)

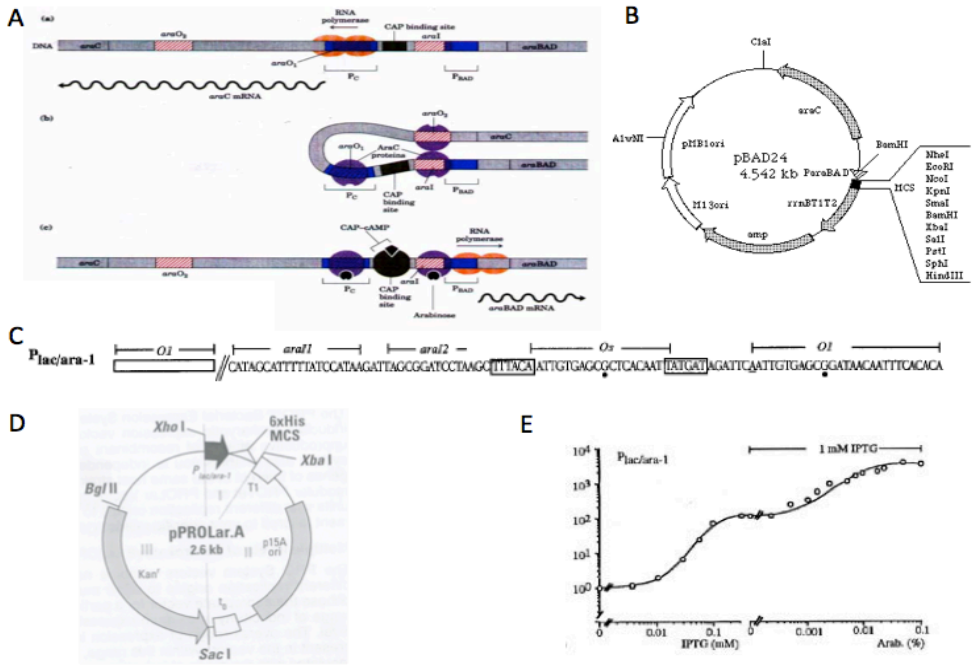

Figure 4.2 (A) The schematic mechanism of pBAD promoter controlled by arabinose; (B) The map of pBAD plasmid used for the expression of toxin genes. The plasmid used in the experiment is chloramphenical resistance (55); (C) The sequence of lac/ara promoter; (D) The map of plasmid pProlar was used for the expression of RNA repair protein; (E) The lac/ara promoter in plasmid pProlar could be induced by both IPTG and arabinose in a two-step mode (56).

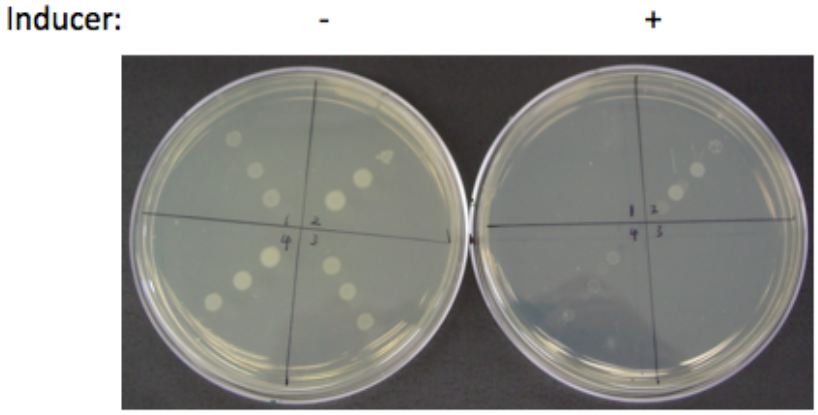

Figure 4.3 The effect of the known toxins on the cell growth (1: colicin D; 2: sarcin; 3: VapC; 4: VapC20). Inducer is 2% arabinose, and it was added into the plate as indicated. The culture of DH5α cells containing indicated plasmids was adjusted to OD600 0.1 and spotted near the center. The adjusted cell culture was subjected to a series of dilution (10 times, 100 times), and spotted towards the rim of plate. Except for sarcin, all the other known toxins inhibit cell growth significantly.

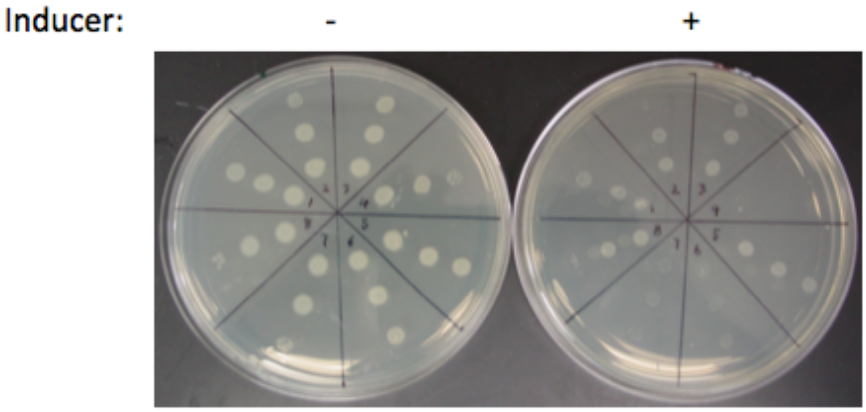

Figure 4.4 The effect of predicted toxins on the cell growth (1: A4YDT6; 2: A4YF78; 3: D0KNK4; 4: *Cgi*-Pin; 5: *Cth*-BECR; 6: *Ec*-EDA; 7: *Pae*-CD, 8: *Str*-CD). The inducer and spotting method is same as Figure 4.3. Predicted toxins 4, 6 and 7 lead to the cell death upon the induction by arabinose.

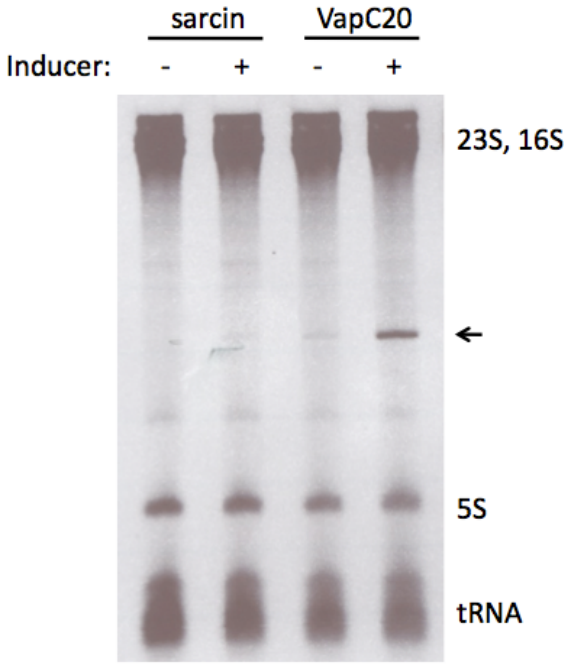

Figure 4.5 Total RNA from the *E.coli* cells. DH5α/P_{BAD}::sarcin and DH5α/ P_{BAD}::SDop::VapC20 (VapC20 plasmid is from Gerdes's group) grow exponentially in LB medium and the expression of toxin genes were induced by the addition of 0.2% arabinose. Cell samples were collected 45 minutes after induction, and the total RNA was extracted and analyzed by 5% PAGE. The arrow indicates the α-fragment caused by the cleavge at the SRL loop of 23S rRNA.

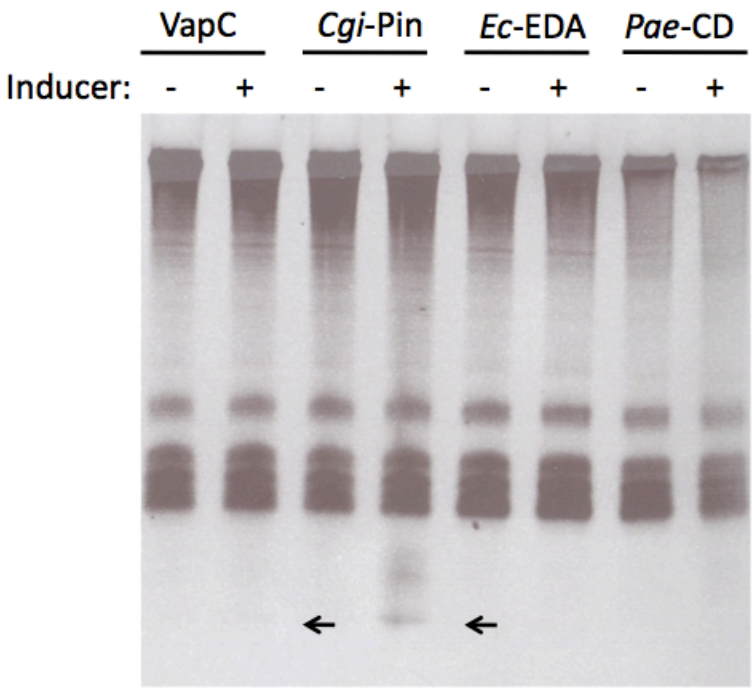

Figure 4.6 Analysis of the total RNA from E.coli cells with the induction of toxin VapC and other three unknown toxins as indicated. The left arrow indicates the cleaved product by VapC, and the right arrow shows the possible cleaved RNA caused by *Cgi*-Pin

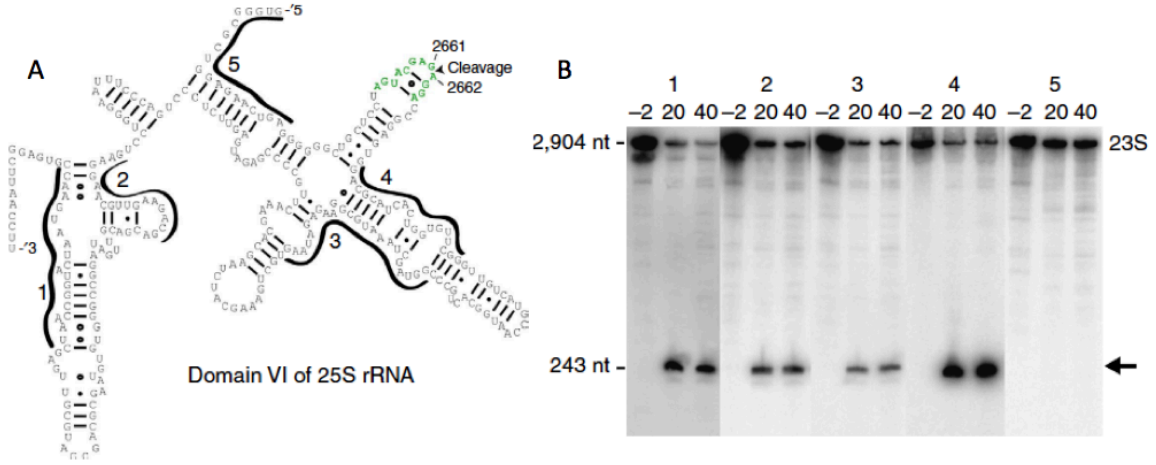

Figure 4.7 Test of α-fragment probe used in Northern blotting (14). (A) The positions of probe candidates are indicated in the secondary structure of domain VI of 23S rRNA. (B) Map the cleavage site by corresponding probes in A. Probe 3 was used for the rest of Northern blotting assays.

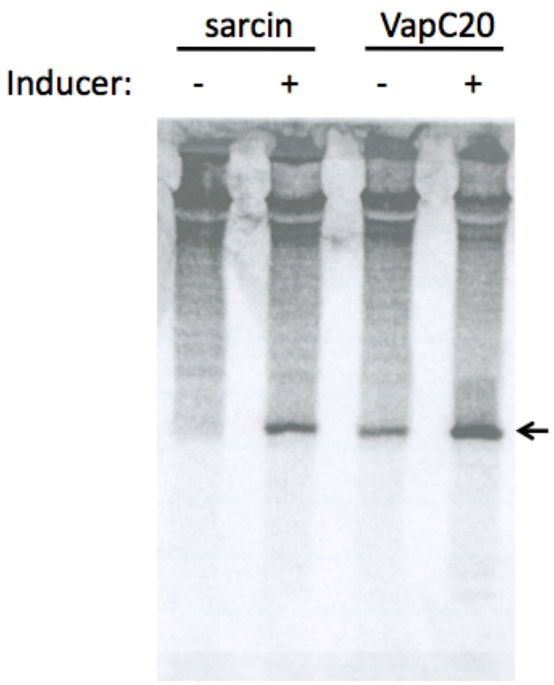

Figure 4.8 Northern blotting of cleaved RNA. Total RNA was prepared and separated in the same way as Figure 4.5, and probe 3 in Figure 4.7 was used. The cleaved product α-fragment is indicated by the arrow.

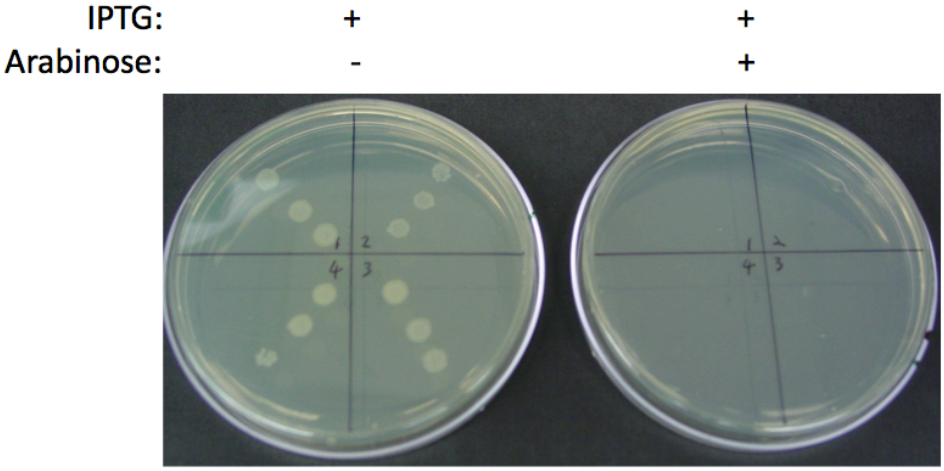

Figure 4.9 Test the recovery of cell growth. pBAD::colicin D and pProlar::repair proteins (1 is empty plasmid, 2 is *Cth*Pnkp/Hen1, 3 is *Ste*Pnkp/Hen1, 4 is *Tde*Pnkp/Hen) were co-transformed into DH5α cells. The cell culture was grown in LB with both chloramphenicol and kanamycin and spotted on the plates the same way as in Figure 4.3. The inducer IPTG (1mM) and arabinose (0.2%) were added into plates as indicated. *Cth*, *Clostridium thermocellum*; *Ste*, *Sebaldella termitidis*; *Tde*, *Treponema denticola*.

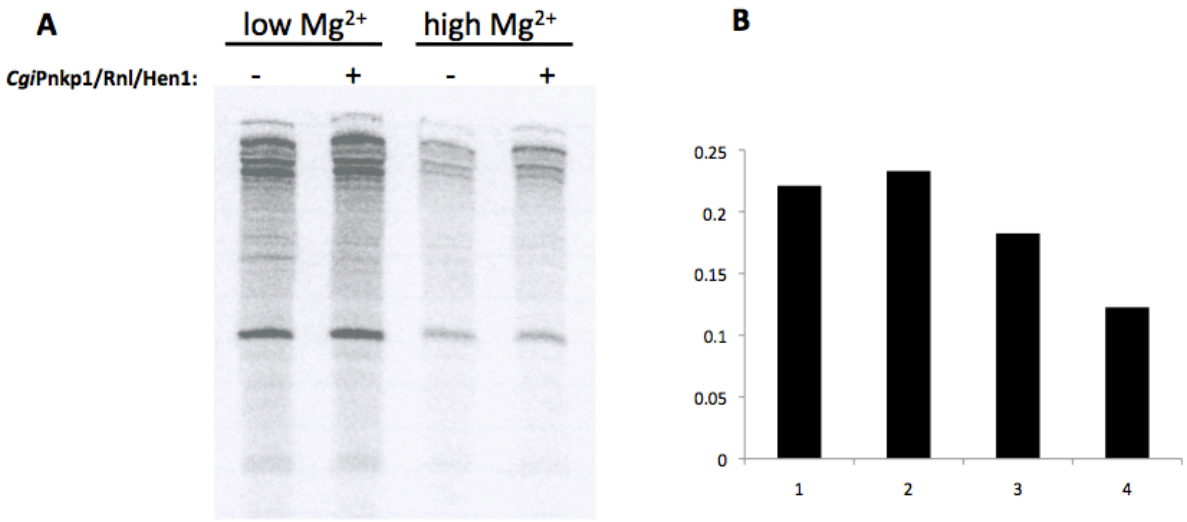

Figure 4.10 Ribosome repair assay. (A) Northern blotting of the cleaved ribosome treated without or with repair protein in the presence of different concentrations of Mg^{2+} (low, 2mM; high, 10 mM). The low Mg^{2+} concentration is reported to force the two subunits of ribosome to fall apart, while a high concentration of Mg^{2+} keep the intact form of ribosome. (B) The quantification of the percentage of cleaved ribosome in A.

CHAPTER 5: STRUCTURAL STUDY OF MAB21D2: A NEWLY IDENTIFIED NTASE MEMBER

1. Introduction

NTase fold proteins constitute a large and highly diverse superfamily. Almost all known members from this superfamily transfer NMP to a hydroxyl group of substrates including protein, nucleic acid and small molecules (Table 1.1). NTase superfamily enzymes catalyze reactions involved in important biological processes, such as DNA repair, RNA editing, antibiotic resistance, signal transduction and so on. The superfamily members are characterized by the presence of common alpha/beta-fold core structure (Figure 1.7), which contains the conserved residues: hG[GS], [DE]h[DE]h and h[DE]h. The three conserved aspartates/glutamates can coordinate divalent ions, which activate the acceptor hydroxyl group of the substrate.

Subgroup XXIV (mab21) is one of the newly identified NTase subfamily. The known members are important for development or signal transduction: Mab21D6 in mice is essential for the development of embryonic brain, eyes, limb and neural crest derivatives; another important member D1 (cGAS) is shown to respond to dsDNA in cytosal and generate cGAMP to activate interferon response through the downstream receptor STING (51-52, 104-107). Remarkably, unlike the cononical cyclic dinucleotides derived from bacteria (53, 108), the cGAMP generated by cGAS contains mixed linkages of both 2'-5' phosphodiester bond and 3'-5' phosphodiester (Cyclic [G(2' -5')pA(3' -5')p]). This noncanonical linkage ensures the opportunity of this second messenger molecule to activate diverse receptor variants (109-111), and may also help to increase its stability since most of RNase target 3'-5' linkage. cGAS is structurally homologus to

another NTase fold protein OAS, which also functions as a innate immune sensor by converting ATP into 2'-5'-linked oligoadenylate upon activation by dsRNA (45, 46, 112)

We studied another member of Mab21 subfamily, Mab21D2, which is closest to cGAS (formerly named Mab21D1). According to sequence alignment, Mab21D2 is highly conserved in vertebrate, and it may possess critical motif of active NTase (Figure 1.10). We crystallized the truncated version of human Mab21D2 and solved the structure (Mab21D2-Δ28) (Figure 5.3). A DALI search revealed that cGAS highly resembled the structure of Mab21D2, even though their sequences were only 30% identical (Figure 5.1). We compared crystal structures of Mab21D2 with cGAS or OAS, especially at the nucleic acid binding cleft and the active site (Figure 5.4, 5.5, 5.6), and found that the apo form of Mab21D2-Δ28 was not fully activated. We also carried out biochemical reactions and analyzed the products though RP-HPLC followed by mass spectrometry and NMR. Based on the data from structural comparison and the biochemical assay, we propose that D2 might exhibit some functions that are different from cGAS.

2. Methods and Materials

2.1 Cloning, overexpression and purification of the recombinant protein

The gene encoding human Mab21D2 was amplified from cDNA library (prepared from Hela cell line) and inserted into the pET-SUMO vector (N-terminal his-SUMO tag) or pET-MBP vector (N-terminal his-MBP tag). Different primers were used to amplify the N-terminal truncation versions of Mb21D2: Δ12, Δ21 and Δ28. The encoding plasmids were transformed into *E. coli BL-21*(DE3) strain. The proteins were all expressed at 18 ºC for 20 hours after induction with 0.5mM IPTG. Cells were harvested by centrifugation and stored at -80 °C.

Cell pellets were thawed, resuspended in lysis buffer (20 mM Tris-HCl, pH 8.0, 10 mM

NaCl, 2% glycerol) and lysed by French press. The cell lysate was centrifuged and the proteins were purified from the supernatant by a FPLC system. The supernatant was first applied to the His- Column, and the fractions contained the target protein was combined and diluted up to 3 times volume with lysis buffer. The proteins with his-SUMO tag were digested by ULP1 (ubiquitin-like protease) protein overnight, while the proteins containing his-MBP tag were digest by TEV protein overnight. After cleavage, the protein was loaded into His column again, and the flow through was collected and further purified by Heparin affinity column and Superdex 200 size exclusion chromatography. All purified proteins were stored in the gel filtration buffer (10 mM HEPES, pH 7.0, 200 mM NaCl, and 1 mM DTT) for further study.

To produce selenomethione-incorporated Mab21D2, *E. coli* Rosetta strain was used for expression and methionine pathway inhibition was used for cell growth. The protein was purified the same way as the wild type described above. Only the protein fused with his-MBP tag is soluble under this growing condition.

2.2 Crystallization, data collection and structure determination

The purified Mb21D2-Δ28 was concentrated to 7 mg/ml, and mixed with a reservoir solution containing 1.3 M $(NH_4)_2SO_4$, 5 mM $MgCl_2$, and 100 mM MES (pH 6.2). The crystal was grown at 18 °C within 3 days by using hanging drop vapor diffusion method. Crystals were soaked in cryoprotecting solution containing all the components of the reservoir solution supplemented with 30% glycerol. The cryo-protected crystals were mounted in nylon loops and flash-frozen in liquid nitrogen. Data were collected at 21-IF beamline at the Advanced Photon Source (APS) and processed by HKL2000. Phase for the structure of the hexamer was determined based on a SAD data from a crystal of the selenium incorporated protein. The initial

model was built automatically with the Phenix software. The remaining model was built using Coot program, followed by refinement using Phenix program. Several rounds of model building, followed by refinement, resulted in a final model of the Mab21D2 with R_{work} and R_{free} of 19.5 % and 25.3%, respectively (67-69).

2.3 RP-HPLC and LC-MS

Double-stranded DNA, and DNA/RNA duplexes were annealed in the annealing buffer containing 70 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 5 mM DTT at equimolar concentrations by heating up to 95 °C for 3 minutes followed by slow decrease to 25 °C. Purified recombinant full length protein was incubated in a 20μl reaction containing 5 mM $MgCl_2$, 150 mM NaCl, 20 mM Tris-HCl, pH 7.5, 1 mM DTT, 10% glycerol and 3 μM indicated type of nucleic acid (54) at 25 °C for overnight. The reaction solutions were applied to 3K centricon, and the flow through was analyzed by RP-HPLC.

RP-HPLC analyses were carried out with a Waters 1525 HPLC system on a C18 Luna column (250 mm x 2 mm, Phenomenex) with a flow rate of 0.3 mL/min. Solvent A was 5 mM ammonium acetate (pH 5.3), solvent B was 40% acetonitrile in water. The analytes were eluted using gradient of 0-50% B in 30 minutes.

ESI-LC-MS experiments were carried out at Mass Spectrometry Center with a Waters 2795 HPLC system and a Waters Q-TOF Ultima API Mass Spectrometer. The absorption was monitored by UV (220-310 nm), and the mass was monitored in both positive and negative modes. The same column, solvents and gradients described above were used for ESI-LC-MS.

2.4 NMR

The fraction corresponding to the new peak in HPLC was collected and dried via SpeedVac system. The dried molecules were dissolved in the 5 mM phosphate buffer in D_2O buffer and analyzed in NMR laboratory.

3. Results and Discussions

3.1. Overall structure of Mab21D2

Mab21D2 is only present in vertebrate, (Figure 1.8), and the proteins from different organisms are highly conserved (Figure 1.9). We cloned the corresponding gene from human Hela cell cDNA library (provided by Xinying Zong), and purified the protein (~56 kDa) to homogeneity. The full-length protein forms dimmer in the slution, and its solubility was relatively low. In order to facilitate overexpression and crystallization, we tested several version of Mab21D2 with N- terminal truncation based on conservation. Deletion of N-terminal 12 or 20 amino acids did not change the behavior of the protein very much, but the protein with N-terminal 28 amino acids deleted exhibited high solubility and behaved as a monomer form in the solution judged by gel filtration chromatography (Figure 5.2).

We successfully crystallized Mab21D2-Δ28, which diffracted to 2.6Å resolution. The phase was determined through a selenomehtionine derivative. After initial model auto building and manual refinement against the native data, we obtained the structure with good statistics (Table 5.1). Since the C-terminal around 60 amino acids could not be determined from the map, the current model only covers residues from 29 to 431.

Similar to cGAS (54, 56), the structure of Mab21D2-Δ28 comprises two lobes (Figure 5.3). Lobe 1 (Figure 5.3, right) contains two twisted β-sheet flanked by two long α-helices (αB,

αC), which possess the NTase fold (Figure 1.7); lobe 2 (left) contains a four-helices bundle (αE-H) and two helices at the both terminus (αA, αI). The two lobes are connected by αA and αB. According to sequence and structure alignment, the essential residues (hG[GS], [DE]h[DE]h and h[DE]h) of the NTase fold superfamily are located at β1, β2, β9 and the loop between the first two β strands.

3.2 Structural comparison of Mab21D2 with other NTases

A DALI search identified cGAS (Mab21D1, PDB: 4lev) as the most closely resembling the fold of Mab21D2 with a z score of 28.4 and RMSD of 2.7 Å. In addition, 2', 5'-oligoadenylate synthase 1 (OAS, PDB: 4ig8) is also structurally similar to Mab21D2, exhibiting a z score of 13.9 and RMSD of 3.4Å.

Structural alignment of D2 and OAS-dsRNA (Figure 5.4A) showed that the NTase core structure (the β-sheet flanked by two helices) aligns better than the rest. Detailed comparison of nucleic acid binding platform and active site reveal that both Mab21D2 and OAS adapt curvature suitable for nucleic acid binding on the oppostion side of active site (Figure 5.4B). The binding of dsRNA is shown to induce structural change of the dsRNA-binding motif in OAS, and make this structurally uncharacterized motif recognize the minor grooves of dsRNA (112). The conformation change in OAS due to dsRNA binding propagates from RNA-protein interface to the active site, resulting in a new helix formed between the first and second β strand (Figure 5.4C). Thus, the catalytic residues would be repositioned in such a way that they could coordinate magnesium ions and bind ATP in the active site. In the structure of Mab21D2, however, the three proposed essential residues are not properly aligned to coordinate the magnesium ions (Figure 5.4D), which might cause the inactivation of Mab21D2. Moreover, the

long loop between the first two β strands of Mab21D2 block the active site (Figre 5.4D). We speculate that upon activation of some nucleic acid, Mab21D2 may also undergo some significant conformation change to re-organize the active site for substrate binding as well as enzyme activation.

Mab21D1 (cGAS) serves as dsDNA sensor in cytosol and generates small molecule cGAMP in order to stimulate downstream immune signaling. As mentioned before, among all known structures, cGAS is the closest to Mab21D2 (Figure 5.5A). We aligned the apo form of cGAS with Mab21D2, and identified similar structural features including the 2 lobes connected by α helix. However, the zinc finger on the nucleic acid binding cleft of cGAS does not exist in D2 (Figure 5.5B). Since zinc finger has been shown to be important for dsDNA binding and enzyme activation of cGAS (56), we propose that Mab21D2 may have different nucleic acid specificity from cGAS. Upon dsDNA binding, cGAS changes its conformation in different places, including the break of long helix (spine), the shift of β-sheet. We aligned the structure of cGAS, cGAS-dsDNA (activated form of cGAS) and Mab21D2 together, and found that Mab21D2 was closer to cGAS-dsDNA structure at the positions of spine and β-sheet (Figure 5.5C, D). Based on theis structural analysis, it appears that Mab21D2 is partially activated with apo enzyme. However, the first two β strands together with the loop in between, which contain the potentially essential catalytic residues, are very different in Mab21D2 when compared to cGAS-dsDNA. A small helix is formed in cGAS-dsDNA to reposition the catalytic residues and widen the reaction pocket, but a long loop in Mab21D2 blocks the potential reacion pocket, presumably preventing substrates from entering the active site (Figure 5.5E).

We also compared the protein surface of Mab21D2 to the activated form of cGAS and OAS, especially near the active site or nucleic acid binding cleft (Figure 5.6). As long as cGAS

and OAS are activated by the nucleic acid through binding, a wide pocket would form with the catalytic residues (Asp/Glu) pointing to the vacant space (Figure 5.6B, C), where the substrates NTPs can bind. However, in the structure Mab21D2, the long loop blocked the proposed active site, making NTP binding impossible. We also observed that there is a vacant space near the long loop in Mab21D2 (Figure 5.6A), and hypothesize that the empty space might accommodate part of loop and help create the reaction pocket as a result of activation, which will presumable create the reaction pocket. The activator binding platforms of cGAS and OAS are predominantly positive charged, which is preferred for binding the negatively charged nucleic acids (Figure 5.6E, F). The corresponding part of Mab21D2 does not exhibit the same degree of positive charge (Figure 5.6D), which may indicate the potential activator of Mab21D2 might not be nucleic acids. Although we have not ruled out the possibility that the nucleic acid could bind to Mab21D2 in an induce-fit mode, it is possible that Mab21D2 works differently from cGAS or OAS. It is even possible that in reverse: it binds a messenger molecular at the active site, which triggers structural change in Mab21D2 that results in signaling through interaction with other factors.

3.3. Test of enzymatic activity of Mab21D2 *in vitro*

We hypothsized that Mab21D2 might also employ nucleotides for enzymatic activity based on its similarity to cGAS (Mab21D1). We first incubated all four kinds of nucleotides with or without Mab21D2, and compared the product through RP- HPLC (Figure 5.7). There were some interesting results based on this preliminary experiment: 1) Among all four nucleotides incubated with the protein, only GTP and ATP were consumed; 2) several new peaks appeared

only in the presence of protein; 3) the new peaks generally appeared after NTPs, indicating that they were less hydrophilic compared to original substrates NTPs.

We further speculated that, as in the case of cGAS, efficient reaction might require the enzymatic activation from binding of nucleic acid. Therefore, so we tested different types of nucleic acids (including ssDNA, ssRNA, dsDNA, DNA/RNA hybrid) by incubating them with Mab21D2 in the reaction only containing GTP and ATP. Based on comparison of commcially-purchased nucleotides (ATP, ADP, AMP, GTP, GDP, GMP), we concluded that most of new peaks generated in the presence of Mab21D2 are the hydrolyzed products of ATP and GTP. The hydrolysis is especially pronounced with the addition of nucleic acids. However, a new peak, which did not match any hydrolyzed product, appeared only when ssDNA or ssRNA was added into the reaction (Figure 5.8D, F, G, labeled "X"). We suspected this the compound of this new peak as the product generated by Mab21D2, and attempted to characterize through further analysis described below.

3.4. Identification of a new product generated by Mab21D2

Since the generation of the new product "X" is accompanied with equal consummation of GTP and ATP, it was reasonable to assume that Mab21D2 acts similarily to cGAS. To identify the product, we employed LC-MS. However, we failed to observe an ion with m/z= 675 (cGAMP) in the positive mode of MS. Further analysis of the sample collected from the new peak "X" also did not match with other dinucleotides any significant signal for other types of dinucleotides (diGMP, diAMP). We tried MS with the negative mode as well, but still failed to identify the new molecule.

We then scaled up the reaction and collected enough compound for NMR analysis. ^{1}H spectrum showed three peaks, which might be from C2, C8 of adeninie, and C8 of guanine (111) (Figure 5.9). We could not find the signal of proton from the ribose based on ^{1}H spectrum, or any obvious signal of phosphatefrom ^{31}P spectrum. At this time, we realized that the new peak "X" might only contain bases instead of nucleotides. This was supported by the fact that the new peak "X" was split into two when another HPLC column was used. With additional experiments, we finally identified the new peak X contain the bases of guanine and hypoxanthine.

4. Figures and Tables

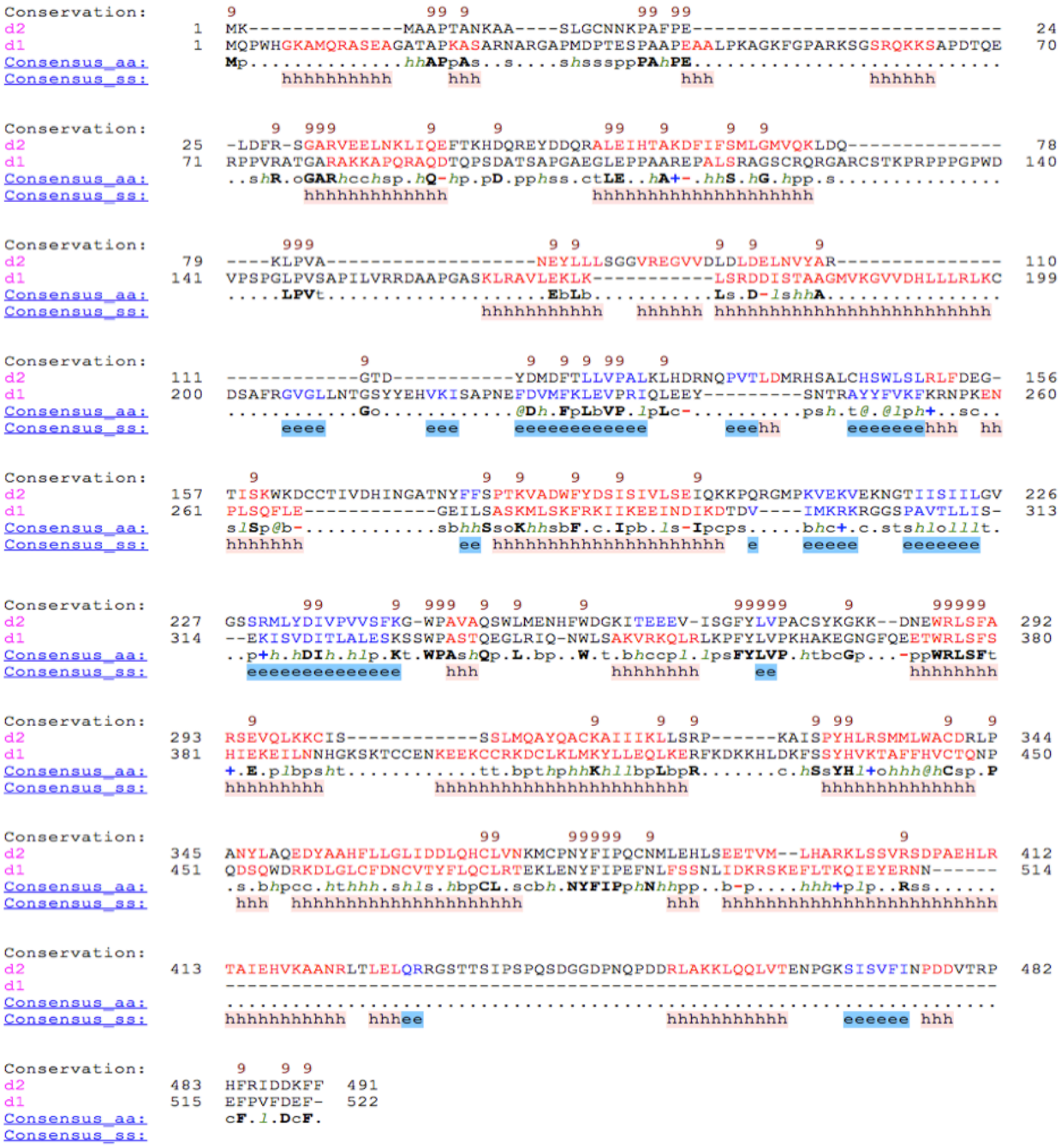

Figure 5.1 Sequence alignment of human Mab21D1 (cGAS) and Mab21D2. The figure is made through Promals3D alignment. The consensus_aa indicated conserved amino acids while consensus_ss showed the conserved structure (h, α helix; e, β strand).

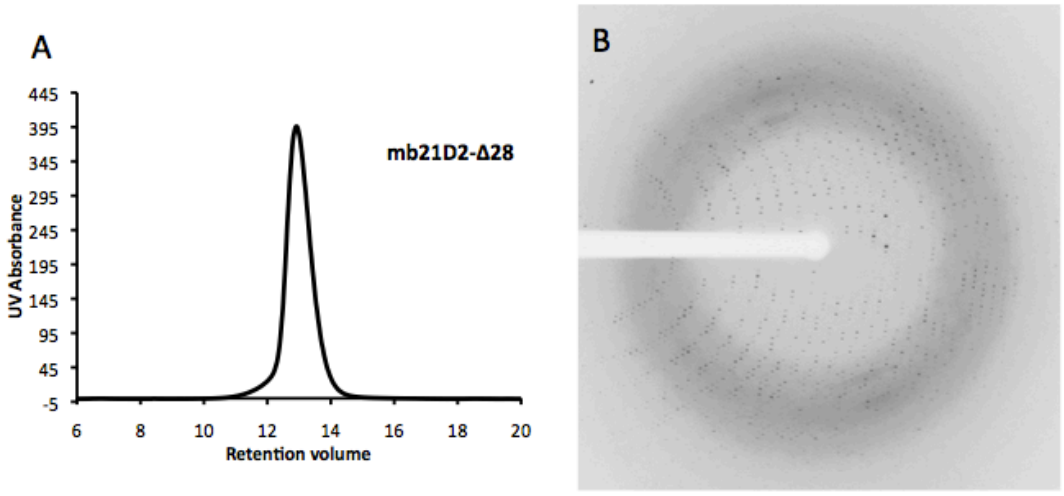

Figure 5.2 The purification and diffraction of Mab21D2-Δ28. (A) The retention volumn of size exclusion chromatography confirms that the protein is monomer in solution. (B) The X-ray diffraction map of the protein shows that the resolution of structure can reach around 2.6Å.The water ring is shown as the black circle.

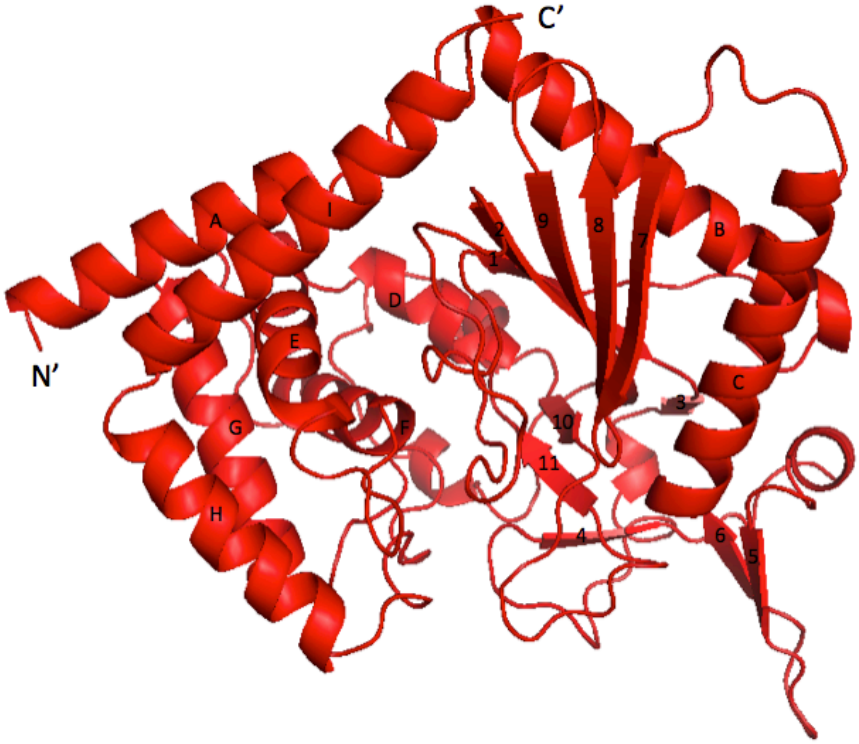

Figure 5.3 Overall structure of Mab21D2-Δ28. The long α helices (A-H) and β strands (1-11) are labeled in the figure, while some short ones are not labeled for the clarity. The structure could only be determined from amino acid 29 to 431 (N', N-terminal; C', C-terminal).

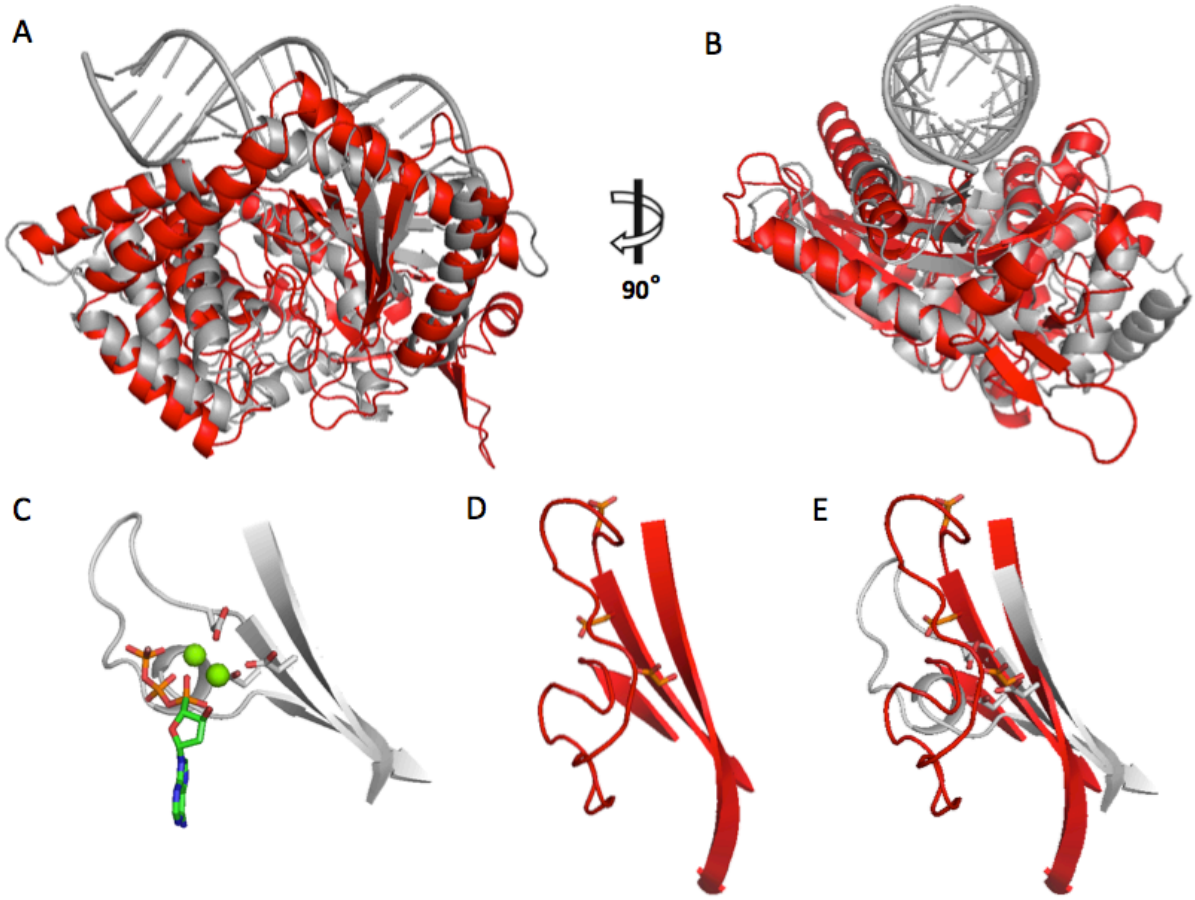

Figure 5.4 Similarity and difference of the structures of Mab21D2 and activated OAS. (A) Structural alignment of Mab21D2 (red) and OAS-dsDNA (gray, PDB: 4ig8) (112). (B) Side view of overall alignment. (C) In the active site of OAS, three catalytic aspartate residues are shown in stick (carbon in white and oxygen in red), which could cooridinate two megnesium ion (two green spheres). The ATP (carbon in green, oxygen in red, phosphorus in organe and nitrogen in blue) shown in stick could be activated by the metal ions. (D) The proposed three catalytic aspartate residues of Mab21D2 are shown in stick (carbons in orange and oxygen in red), and they are located in the β-sheet as well as the loop connecting the first two β strands. (E) The alignment of C and D shows that the aspartate residues in Mab21D2 are not well arranged for the reaction. Moreover, there is a long loop in Mab21D2 structure blocks the reaction pocket.

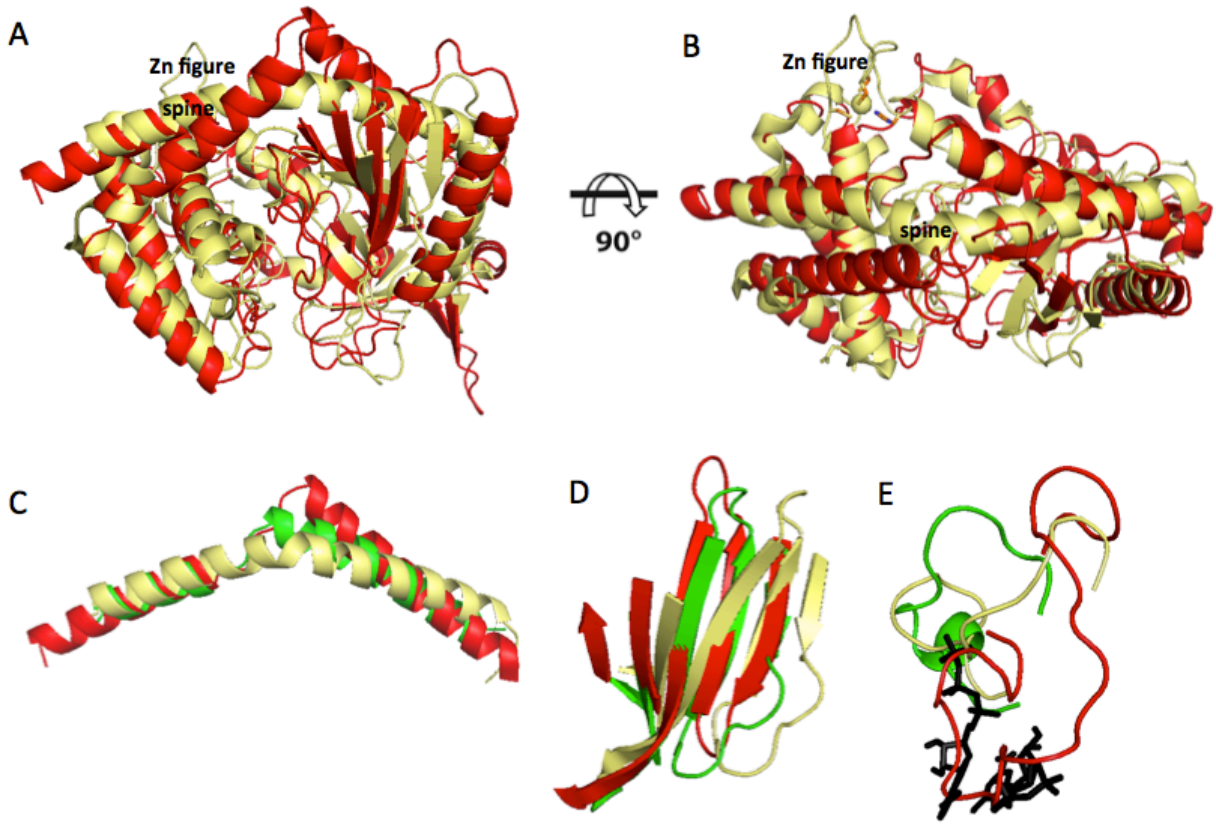

Figure 5.5 Similarity and difference of the structures of Mab21D2 and cGAS. (A) Structural alignment of Mab21D2 (red) and cGAS (yellow, the structure element spine and Zn finger are indicated in the figure, PDB: 4k8v) (54). (B) Top view of overall alignment. (C) The alignment of the spine in Mab21D2 (red), cGAS-dsDNA (green, PDB: 4k96) and cGAS (yellow) shows that the spine part is broken into two helices in the first two structures, but not in cGAS structure. (D) The alignment of the NTase core fold β-sheet in Mab21D2, cGAS-dsDNA and cGAS (color in the same way as C) indicates that the β-sheet of cGAS undergoes shift upon dsDNA binding, and the β-sheet in Mab21D2 is close to that in activated form of cGAS. (E) The alignment of the loop between β1 and β2 strand in three structures. A short helix is induced when cGAS is bound to dsDNA (compare green with yellow), making the pocket accessible to the substrates ATP and GTP (shown in black stick). The loop in Mab21D2 (red) occupies the substrate-binding site.

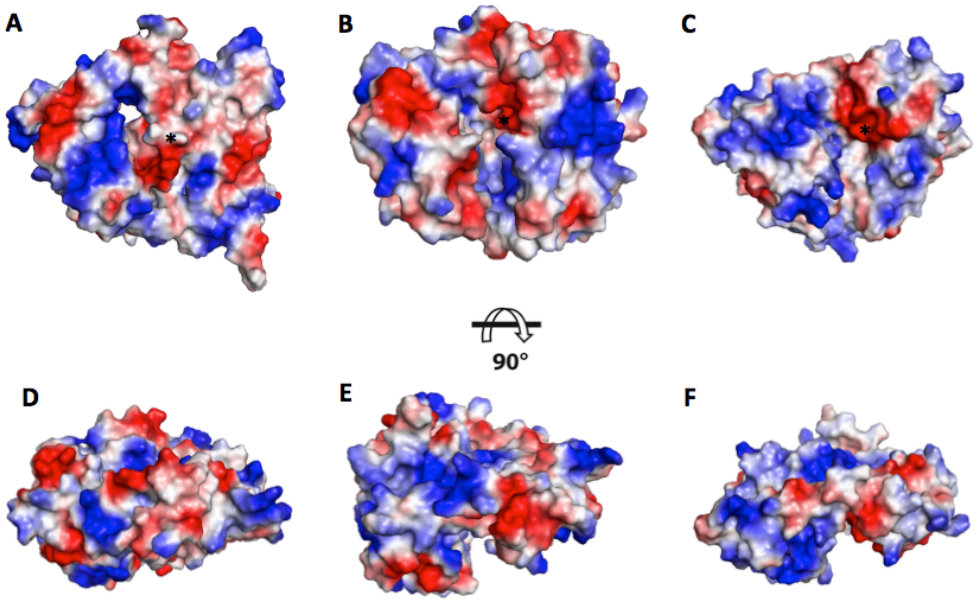

Figure 5.6 Surfaces of the proteins are colored according to electrostatic potential (blue, basic; white, neutral; red, acidic). Mab21D2 (A), activated form of cGAS (B) (PDB: 4k96) and OAS (C) (PDB: 4ig8) are in the same orientation as Figure 5.4A, 5.5A (active sites or potential catalytic site are indicated by asterisk). (C, D, F) Top view of A, B, C shows the platform for the binding of nucleic acid.

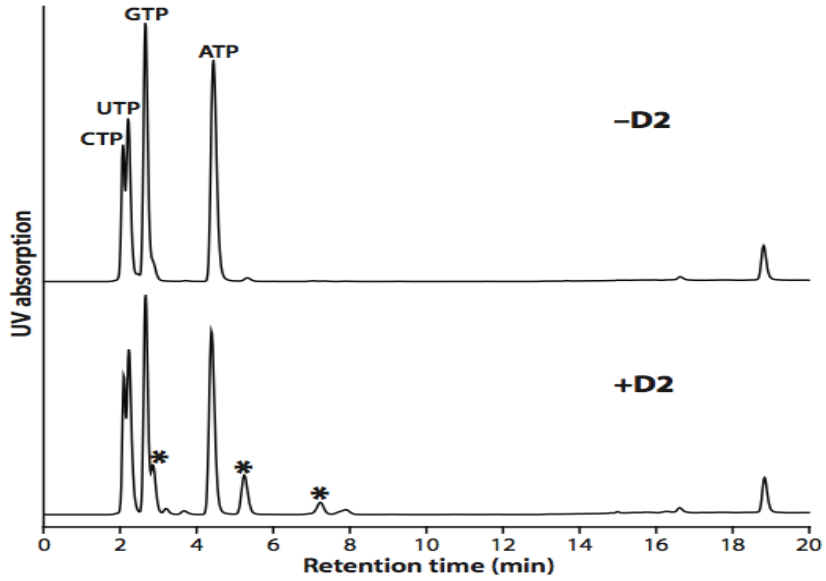

Figure 5.7 HPLC analysis of enzymatic function of full-length Mab21D2. Four different types of NTPs (ATP, GTP, UTP, CTP) were incubated without (top panel) or with Mab21D2 protein (bottom panel), and the small molecules were analyzed by HPLC. The asterisks indicate the new peaks formed in the presence of Mab21D2 protein. The two panels are of the same scale; thus the peaks corresponding to UTP and CTP barely change with the addition of protein.

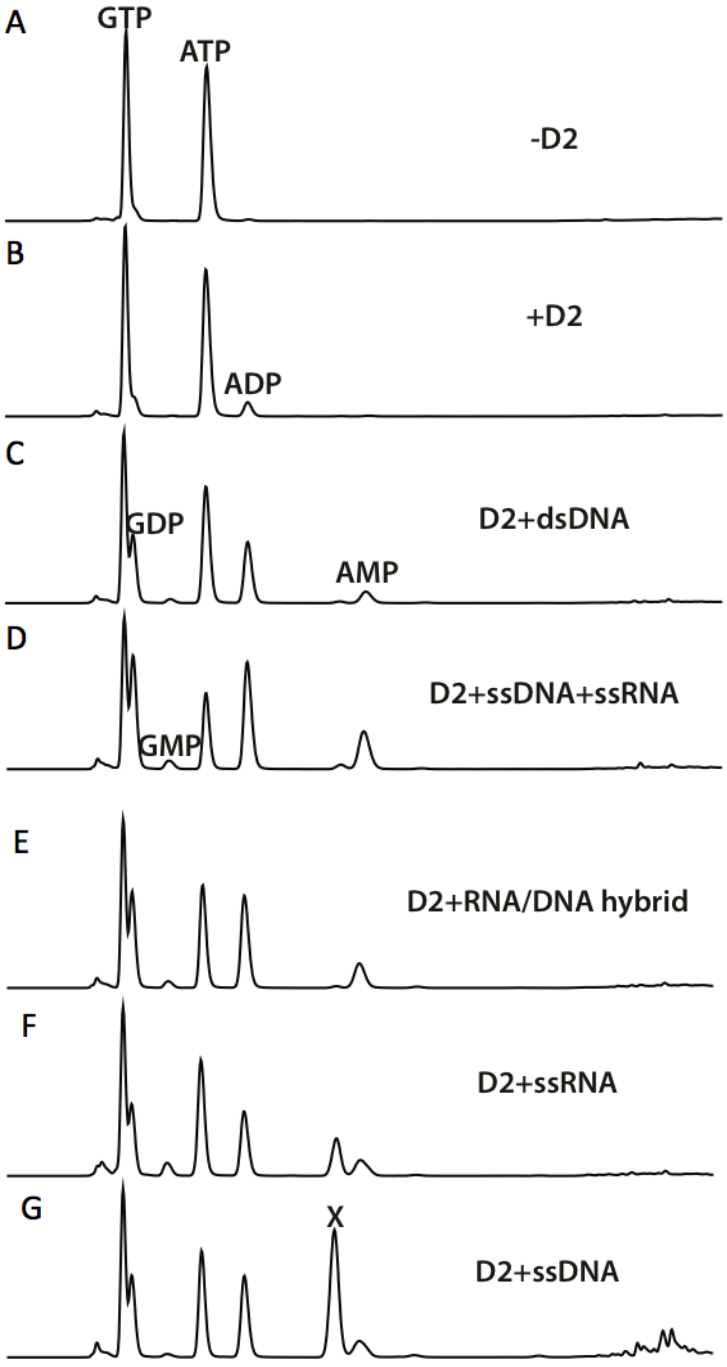

Figure 5.8 HPLC analysis of nucleic acid activator for Mab21D2. GTP and ATP were incubated with protein Mab21D2 and nucleic acid as indicated (A-G), and the small molecules were analyzed by HPLC. The hydrolysis products (GDP, GMP, ADP, AMP) are labeled according to the standard. "X" indicates the potential product with unknown identity. (ss, single strand; ds, double strand; RNA/DNA hybrid, a RNA strand and a DNA strand containing complementary sequences; ssDNA+ssRNA, a RNA strand and a DNA strand sharing the same sequence).

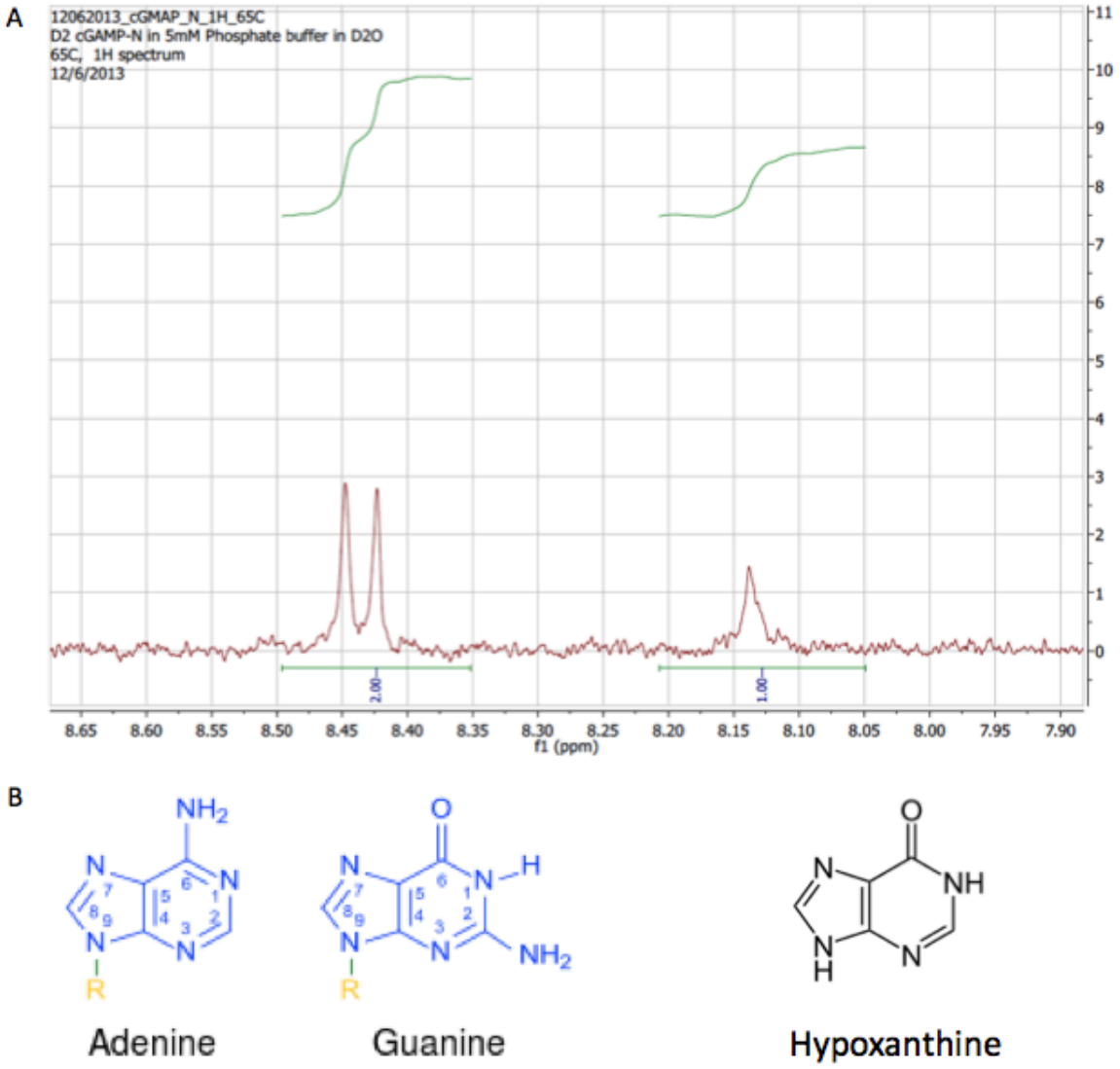

Figure 5.9 NMR analysis of the new product. (A) The ^{1}H NMR characterization of the sample collected from peak X (Figure 5.7G) is as follows: ^{1}H NMR (500 MHz, 5 mM $NaH_2PO_4/Na_2HPO4/D_2O$, c = 1 mM, 65°C) δ 8.45 (s, 1H), 8.42(s, 1H), 8.14 (s, 1H). (B) The chemical stucture of the bases involved in this study. The protons at C2, C8 of adenine (hypoxanthine is similar to adenine) and C8 of guanine would show the chemical shift in ^{1}H NMR spectrum.

Table 5.1 Data collection and refinement statistics for the crystal in chapter 5

	Mab21D2-Δ28
Data collection	
Data type	SAD/Native
Space Group	*I*222
Cell dimensions	
a, b, c (Å)	102.7, 126.7, 198.2
α, β, γ (°)	90.0, 90.0, 90.0
Resolution (Å)	42.1-2.6
Wavelength (Å)	0.9787
Redundancy	5.6
Completeness (%)	98.9 (92.5)
$I/\sigma I$	12.4
R_{merge} (%)	5.3
No. reflections	39,599 (3,667)
R_{work}/R_{free} (%)	19.5/25.3 (22.8/29.6)
No. atoms	6,430
Protein	6,430
Ligand/ion	0
Water	0
B factors (Å^2)	
Protein	53.0
Ligand/ion	0
Water	0
r.m.s. deviations	
Bond lengths (Å)	0.009
Bond angles (°)	1.24

*Values in parenthesis are for highest resolution shell
Same dataset is used for phase determination and structure refinement

CHAPTER 6: CONCLUSIONS AND SUMMARY

1. Molecular basis of bacterial Hen1 activating the ligase activity of bacterial Pnkp

RNA damage caused by most ribotoxins can result in broken RNAs with 2', 3'-cyclic phosphate group at 3'end and 5'-OH at 5'end. It has been shown that the genes encoding bacterial Pnkp and Hen1 appear pair-wise in the same operon, and the two proteins can form a complex to carry out RNA repair including RNA healing (remove the phosphate at 3'-end and phosphorylate 5'-OH) and sealing process (ligate the processed RNA) (Figure 2.2) (34). Bacterial Pnkp alone can heal the broken RNA, but it alone cannot seal the broken RNA, even though it contains all the essential motifs for a RNA ligase (30).

We crystallized a functional ligase consisting of the C-terminal half of Pnkp (Pnkp-C) and the N-terminal half of Hen1 (Hen1-N) from *Clostridium thermocellum* (Figures 2.5, 2.6). The structure reveals that the N-terminal domain of *Cth*Hen1, shaped like a left hand, grabs the flexible insertion module of *Cth*Pnkp and locks its conformation via further interaction with the C-terminal addition module of *Cth*Pnkp. Formation of the *CthPnkp*-C/Hen1-N heterodimer creates a ligation pocket with a width for two strands of RNA, depth for two nucleotides, and the adenosine monophosphate (AMP)-binding pocket at the bottom (Figure 2.11) (58).

RNA ligation usually contains three steps: 1) the adenylylation of enzyme, 2) the adenylylation of 3'-half RNA and 3) the ligation of two halves of RNA. The structure of *Cth*Pnkp-C/AMP solved by Shuman's group indicates that Pnkp alone was able to finish the step-1 reaction (Figure 2.12). We carried out the step-2 specific assay, which clearly showed that Hen1 could greatly enhance the transfer of AMP from enzyme to RNA (Figure 2.13). We concluded that the activation of ligase activity by Hen1 starts at the step-2 of ligation. Mutation studies reveal that efficient RNA ligation requires direct involvement of several conserved

residues from the insertion module as well as Hen1-N (Figure 2.14). In addition to playing an essential structural role in the formation of a ligation pocket for RNA ligation, Hen1-N may also provide its own residues that might directly participate in RNA ligation.

The requirement of Hen1 in the ligation process ensures the opportunity of 2'-O-methylation carried out by Hen1 C-terminal methyltransferase during RNA repair. Due to the lack of 2'-OH group at the repair junction, the repaired product is resistant to the cleavage caused by the same toxins (34).

2. Reconstitution and structure of a new bacterial RNA repair complex Pnkp1/Rnl/Hen1

To investigate whether distant Hen1 is present in bacteria, we performed comprehensive BLASTP search using the sequence of *Ava*Hen1 belonging to the Pnkp/Hen1 system. We discovered ten bacterial species with a gene encoding Hen1 that is modestly homologous (~20%) to *Ava*Hen1. However, unlike Pnkp/Hen1 system, a gene encoding Pnkp homolog could not be found near *hen1* in these organisms. Instead, an operon that is far away from *hen1* was found to contain two genes encoding proteins that may be involved in RNA repair: one encodes a protein that is equivalent to the ligase domain (Rnl) of Pnkp with modest sequence identities (~20%) and the other encodes a protein (Pnkp1) homologous (~30% identities) to *T4*Pnkp (Figure 3.2).

We cloned genes encoding Pnkp1, Rnl, and Hen1 from *Capnocytophaga gingivalis* (*Cgi*), and purified the recombinant proteins for analysis. We found that these three proteins could form a heterohexamer, which is functionally equivalent to the bacterial Pnkp/Hen1 heterotetramer (Figure 3.3). *In vitro* reconstitution shows that efficient repair of ribotoxin-cleaved tRNA only occurs in the presence of all three proteins (Figure 3.4, 3.5). Based on kinetics assays, we found that hexamer exhibits higher efficiency than the corresponding hexamer component in terms of

kinase, phosphatase and methyltransferase activities. Moreover, the RNA ligation is more efficient in the presence of SAM (methyl group donor) (Figure 3.6). We conclud that the hexamer is likely the functional unit *in vivo* and this repair complex prefers to generate a methylated repair product.

We crystallized the Pnkp1/Rnl/Hen1 heterohexamer and solved the structure at 3.3 Å resolution. The crystal structure of the complex reveals the molecular architecture of the heterohexamer as two rhomboid-shaped ring structures of Pnkp1/Rnl/Hen1 heterotrimer fused at the Pnkp1 dimer interface (Figure 3.8). The four active sites required for RNA repair are located on the inner rim of each ring but contributed *in trans*, suggesting a novel mechanism of processive RNA repair with efficient 2′-O-methylation (Figure 3.16).

The majority of bacteria possessing Pnkp1/Rnl/Hen1 repair complex were found in gingival plagues in the human mouth, providing a possibility of treating periodontal and dental diseases with small molecules that inhibit RNA repair (Figure 3.18).

3. *In vivo* study of RNA damage and RNA repair

The structural and biochemical studies of the RNA repair complexes help to answer how bacterial RNA repair systems ensure opportunity and efficiency of 2′-O-methylation during repair process, but little is known about the *in vivo* biological function of RNA repair. The previous assays showed the RNA repair systems do not exhibit obvious substrate specificity, so we considered it as the generic RNA repair machinery. In order to mimic cellular response under stress, we introduced the toxin and RNA repair genes into bacteria and regulated their expression (55, 56) in order to provide insight into their *in vivo* function

We confirmed the lethal toxicity of some known toxins by testing the cell growth upon

induction of toxin gene expression. We found that the toxin sarcin does not kill the bacteria efficiently since it has been reported to cleave eukaryotic ribosomes at the SRL region (Figure 4.3). However the cleavage still occurs in bacterial cells since the cleaved fragment was observed among total RNAs as well as in the northern blotting. Moreover, the fragment generated by sarcin disappears over the time, which explains the survival of cells. By using this method, we are able to identify some new toxins (Figure 4.4).

Next, we turned on both toxin and repair genes in cells, but failed to see the recovery of cellular growth (Figure 4.9). It is possible that relative doses of proteins need to be fine-tuned and optimized in the future studies.

4. Structural and functional studies of Mab21D2

Almost all known members in the large and highly diverse NTase fold superfamily transfer NMP to a hydroxyl group of substrate including proteins, nucleic acids and small molecules (42). A new member of NTase named cGAS was reported to respond to dsDNA in cytosol and generate the second messenger cGAMP to trigger the downstream immune response (51, 52).

We solved the structure of human Mab21D2 belonging to the same subfamily as cGAS (Figure 5.3). The structure comprises two lobes connected by two helices, and one lobe possesses the NTase core fold. The overall structure of Mab21D2 highly resembles cGAS, especially in the NTase core fold part. Detailed comparison showed that the structure of Mab21D2 is between the apo and the activated forms of cGAS (Figure 5.5). Unlike cGAS that binds dsDNA, the surface charge of Mab21D2 does not support strong binding of nucleic acids (Figure 5.6). Moreover, the conserved catalytic residues are not well aligned between cGAS and

Mab21D2, indicating that Mab21D2 may not be enzymatic active or it might carry out a reaction that is differenct from cGAS.

In order to find the cognate activator of Mab21D2, we tried different types of nucleic acids. We observed hydrolysis of NTP and generation of some unknown molecules, which were later identified as a mixture of guanine and hypoxanthine (Figures 5.8, 5.9). At present, we are not able to reveal the biochemical function of Mab21D2. The following are two hypotheses, which require future work: 1) the cognate activators might bind Mab21D2 and cause the conformational change of the flexible activation loop to reposition the catalytic residues; 2) Mab21D2 might serve as a receptor of a second messenger molecule, such as c-di-GMP, in order to trigger immune system in response to bacteria infection.

REFERENCES

1. Alfonso Bellacosa, Eric G. Moss (2003) RNA Repair: Damage Control. Current Biology, 13: R482–R484.

2. Jonathan Houseley, David Tollervey (2009) The Many Pathways of RNA Degradation. Cell, 136: 763–776.

3.Douglas L. Black. (2003) Mechanisms of Alternative Pre-Messenger RNA Splicing. Annual Review of Biochemistry 72 (1): 291–336.

4. Suzanne Clancy. (2008) RNA Splicing: Introns, Exons and Spliceosome. Nature Education 1 (1): 31.

5. Christopher R. Trotta, *et al*. (1997) The Yeast tRNA Splicing Endonuclease: A Tetrameric Enzyme with Two Active Site Subunits Homologous to the Archaeal tRNA Endonucleases. Cell 89 (6): 849–858.

6.Shawn K. Westaway, *et al*. (1988) Structure and function of the yeast tRNA ligase gene. The Journal of biological chemistry 263 (7): 3171–6.

7. John Abelson, et al. (1998) tRNA Splicing. Journal of Biological Chemistry 273 (21): 12685–12688

8. Robert W. Hartley, John R. Smeaton. (1973) On the reaction between the extracellular ribonuclease of bacillus amyloliquefaciens (Barnase) and its intracellular inhibitor (barstar). J. Biol. Chem., 248: 5624-5626

9. Pedersen K, et al. (2003) The bacterial toxin RelE displays codon-specific cleavage of mRNAs in the ribosomal A site. Cell 112:131–140.

10. Neubauer C, *et al*. (2009) The structural basis for mRNA recognition and cleavage by the ribosome-dependent endonuclease RelE. Cell 139:1084–1095.

11. Endo Y, *et al*. (1988) The cytotoxins alpha-sarcin and ricin retain their specificity when tested on a synthetic oligoribonucleotide (35-mer) that mimics a region of 28 S ribosomal ribonucleic acid.J Biol Chem 263(17): 7917-20.

12. Endo Y, Mitsui K, Motizuki M, Tsurugi K (1987) The mechanism of action of ricin and related toxic lectins on eukaryotic ribosomes. The site and the characteristics of the modification in 28 S ribosomal RNA caused by the toxins. J Biol Chem 262: 5908–5912.

13. Endo Y, Wool IG (1982) The site of action of alpha-sarcin on eukaryotic ribosomes. The sequence at the alpha-sarcin cleavage site in 28 S ribosomal ribonucleic acid. J Biol Chem 257: 9054–9060.

14. Kristoffer S. Winther, *et al*. (2013) VapC20 of Mycobacterium tuberculosis cleaves the Sarcin–Ricin loop of 23S rRNA. Nature communication 4:2796

15. J. Konisky, *et al*. (1982) Colicins and other bacteriocins with established modes of action, Ann, Rev. Microbiol. 36: 125–144.

16. B.W. Senior, I.B. Holland (1971) Effect of colicin E3 upon the 30S ribosomal subunit of Escherichia coli, Proc. Natl. Acad. Sci. USA. 68: 959–963.

17. C.M. Bowman, J. Sidikaro, M. Nomura. (1971) Specific inactivation of ribosomes by colicin E3 in vitro and mechanism of immunity in colicinogenic cells, Nature New Biol. 234: 133–137.

18. Blanga-Kanfi, S., Amitsur, M., Azem, A., and Kaufmann, G. (2006) PrrC-anticodon nuclease: functional organization of a prototypal bacterial restriction RNase. Nucleic Acids Res. 34: 3209–3219.

19. Jiang, Y., Blanga, S., et al. (2002) Structural features of $tRNA^{Lys}$ favored by anticodon nuclease as inferred from reactivities of anticodon stem and loop substrate analogs. J Biol Chem 277: 3836-3841.

20. Ogawa T, *et al*. (1999) A cytotoxic ribonuclease targeting specific transfer RNA anticodons. Science 283:2097–2100.

21. Tomita K, Ogawa T, Uozumi T, Watanabe K, Masaki H. (2000) A cytotoxic ribonuclease which specifically cleaves four isoaccepting arginine tRNAs at their anticodon loops. Proc Natl Acad Sci USA 97:8278–8283.

22. Haruhiko Masaki , Tetsuhiro Ogawa. (2002) The modes of action of colicins E5 and D, and related cytotoxic tRNases. Biochimie 84: 433–438

23. Winther KS, Gerdes K. (2011) Enteric virulence associated protein VapC inhibits translation by cleavage of initiator tRNA. Proc Natl Acad Sci USA 108:7403–7407.

24. Eric M Phizicky. (2008) Waste not, want not: a case for tRNA repair. Nature structural & molecular biology, 15: 783-784

25. Ougland R, *et al*. (2004) AlkB restores the biological function of mRNA and tRNA inactivated by chemical methylation. Mol Cell. 16(1):107-16.

26. Zhu L, Deutscher MP. (1987) tRNA nucleotidyltransferase is not essential for Escherichia coli viability. EMBO J. 6(8): 2473-7.

27. Amitsur, M., Levitz, R., and Kaufman, G. (1987) Bacteriophage T4 anticodon nuclease, polynucleotide kinase, and RNA ligase reprocess the host lysine tRNA. EMBO J. 6: 2499–2503.

28. Bullard, D.R., and Bowater, R.P. (2006) Direct comparison of nick-joining activity of the nucleic acid ligases from bacteriophage T4. Biochem. J. 398: 135–144.

29. El Omari, K., *et al*. (2005) Molecular architecture and ligand recognition determinants for T4 RNA ligase. J. Biol. Chem. 281: 1573–1579.

30. Martins A, Shuman S. (2005) An end-healing enzyme from Clostridium thermocellum with 5'kinase, 2',3'phosphatase, and adenylyltransferase activities. RNA 11: 1271–1280.

31. Smith P, Wang LK, Nair PA, Shuman S. (2012) The adenylyltransferase domain of bacterial Pnkp defines a unique RNA ligase family. Proc Natl Acad Sci USA 109: 2296–2301.

32. Nandakumar J, Schwer B, Schaffrath R, Shuman S. (2008) RNA repair: An antidote to cytotoxic eukaryal RNA damage. Mol Cell 31: 278–286.

33. Jain R, Shuman S. (2010) Bacterial Hen1 is a 3'-terminal RNA ribose 2'-O-methyltransferase component of a bacterial RNA repair cassette. RNA 16: 316–323

34. Chio Mui Chan, Zhou Chun, Raven H. Huang. (2009) Reconstituting Bacterial RNA Repair and Modification in Vitro. Science, 326: 247

35. Chio Mui Chan, Zhou Chun, Joseph Brunzelle, Raven H. Huang. (2009) Structural and biochemical insights into 2'-O-methylation at the 3'-terminal nucleotide of RNA by Hen1. Proc Natl Acad Sci USA, 106: 17699-17704

36. Huang RH. (2012) Unique 2'-O-methylation by Hen1 in eukaryotic RNA interference and bacterial RNA repair. Biochemistry. 51 (20):4087-95.

37. Saito K, *et al*. (2007) Pimet, the Drosophila homolog of HEN1, mediates 2'-O-methylation of Piwi- interacting RNAs at their 3'-ends. Genes Dev 21: 1603–1608.

38. Yu B, *et al*. (2005) Methylation as a crucial step in plant microRNA biogenesis. Science, 307: 932–935.

39. Li J, Yang Z, Yu B, Liu J, Chen X. (2005) Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in Arabidopsis. Curr Biol, 15:1501–1507.

40. Kurth HM, Mochizuki K. (2009) 2'-O-methylation stabilizes Piwi-associated small RNAs and ensures DNA elimination in Tetrahymena. RNA, 15: 675–685.

41. Zhang C, Chan CM, Wang P, Huang RH. (2012). Probing the substrate specificity of the bacterial Pnkp/Hen1 RNA repair system using synthetic RNAs. RNA, 18: 335-44.

42. Krzysztof Kuchta, et al. (2009) Comprehensive classification of nucleotidyltransferase fold proteins: identification of novel families and their representatives in human. Nucleic Acids Research, 37: 7701–7714

43. Garcia-Diaz, M., et, al. (2005) A closed conformation for the Pol lambda catalytic cycle. Nat. Struct. Mol. Biol., 12: 97–98.

44. Pedersen, L.C., Benning,M.M. and Holden,H.M. (1995) Structural investigation of the antibiotic and ATP-binding sites in kanamycin nucleotidyltransferase. Biochemistry, 34: 13305–13311.

45. Hartmann,R., Justesen,J., Sarkar,S.N., Sen,G.C. and Yee,V.C. (2003) Crystal structure of the 2'-specific and double-stranded RNA-activated interferon-induced antiviral protein 2'-5'-oligoadenylate synthetase. Mol. Cell, 12: 1173–1185.

46. Rogozin,I.B., Aravind,L. and Koonin,E.V. (2003) Differential action of natural selection on the N and C-terminal domains of 2'-5' oligoadenylate synthetases and the potential nuclease function of the C-terminal domain. J. Mol. Biol., 326: 1449–1461.

47. Chow KL, Hall DH, Emmons SW. (1995) The mab-21 gene of Caenorhabditis elegans encodes a novel protein required for choice of alternate cell fates. Development, 121: 3615–3626.

48. Nikolas Nikolaidis, *et al*. (2007) Ancient Origin of the New Developmental Superfamily DANGER. PLoS ONE 2(2): e204.

49. Wong YM, Chow KL. (2002) Expression of zebrafish mab21 genes marks the differentiating eye, midbrain and neural tube. Mech Dev 113: 149–152.

50. Mariani M, *et al*. (1999) Two murine and human homologs of mab-21, a cell fate determination gene involved in Caenorhabditis elegans neural development. Hum Mol Genet, 8: 2397–2406.

51. Sun, L. *et al*. (2013) Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. Science, 339: 786–791.

52. Wu, J., *et al*. (2013). Cyclic GMP-AMP is an endogenous second messenger in innate immune signaling by cytosolic DNA. Science, 339: 826–830.

53. Yin, Q., *et al*. (2012). Cyclic di-GMP sensing via the innate immune signaling protein STING. Mol. Cell 46, 735–745.

54. Pu Gao, *et al*. (2013) Cyclic [G(2',5')pA(3',5')p] Is the Metazoan Second Messenger Produced by DNA-Activated Cyclic GMP-AMP Synthase. Cell, 153: 1–14

55. Xin Li, *et al*. (2013) Cyclic GMP-AMP Synthase Is Activated by Double-Stranded DNA-Induced Oligomerization. Immunity, 39: 1019–1031

56. Filiz Civril, *et al*. (2013) Structural mechanism of cytosolic DNA sensing by cGAS. Nature, 498, 332-337

57. Junzo Otera. (1993) Transesterification. Chem. Rev, 93: 1449-1470

58. Wang P, *et al*. (2012) Molecular basis of bacterial protein Hen1 activating the ligase activity of bacterial protein Pnkp for RNA repair. Proc Natl Acad Sci USA. 109: 13248-53.

59. Shuman S, Lima CD. (2004) The polynucleotide ligase and RNA capping enzyme superfamily of covalent nucleotidyltransferases. Curr Opin Struct Biol, 14: 757–764.

60. Ho CK, Wang LK, Lima CD, Shuman S. (2004) Structure and mechanism of RNA ligase. Structure, 12: 327–339.

61. El Omari K, *et al*. (2005) Molecular architecture and ligand recognition determinants for T4 RNA ligase. J Biol Chem, 281: 1573–1579.

62. Nandakumar J, Shuman S, Lima C.D (2006) RNA ligase structures reveal the basis for RNA specificity and conformational changes that drive ligation forward. Cell, 127: 71–84.

63. Pascal JM, O'Brien PJ, Tomkinson AE, Ellenberger T. (2004) Human DNA ligase I completely encircles and partially unwinds nicked DNA. Nature, 432: 473–478.

64. Lehman IR. (1974) DNA ligase: Structure, mechanism, and function. Science, 186: 790–797.

65. Torchia C, Takagi Y, Ho CK. (2008) Archaeal RNA ligase is a homodimeric protein that catalyzes intramolecular ligation of single-stranded RNA and DNA. Nucleic Acids Res, 36: 6218–6227.

67. Holm L, Kääriäinen S, Rosenström P, Schenkel A. (2008) Searching protein structure databases with DaliLite v.3. Bioinformatics, 24: 2780–2781.

68. Adams, P. D. *et al*. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr, 66: 213-221.

69. Emsley, P. & Cowtan, K. (2004) Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr, 60: 2126-2132.

70. Galburt, E. A., Pelletier, J., Wilson, G. & Stoddard, B. L. (2002) Structure of a tRNA repair enzyme and molecular biology workhorse: T4 polynucleotide kinase. Structure, 10: 1249-1260

71. Eastberg, J. H., Pelletier, J. & Stoddard, B. L. (2004) Recognition of DNA substrates by T4 bacteriophage polynucleotide kinase. Nucleic Acids Res, 32: 653-660

72. Li Kai Wang, Chiristopher D. Lima, Stewart Shuman. (2002) Structure and mechanism of T4 polynucleotide kinase: an RNA repair enzyme. The EMBO Journal, 21: 3873-3880

73. Zhu, H., Smith, P., Wang, L. K. & Shuman, S. (2007) Structure-function analysis of the 3' phosphatase component of T4 polynucleotide kinase/phosphatase. Virology, 366: 126-136

74. Garces, F., Pearl, L. H. & Oliver, A. W. (2011) The structural basis for substrate recognition by mammalian polynucleotide kinase 3' phosphatase. Mol Cell. 44: 385-396.

75. Ushati Das and Stewart Shuman. (2013) Mechanism of RNA 2',3'-cyclic phosphate end healing by T4 polynucleotide kinase–phosphatase. Nucleic Acids Research, 41: 355–365

76. Niroshika Keppetipola, Stewart Shuman. (2006) Mechanism of the phosphatase component of Clostridium thermocellum polynucleotide kinase-phosphatase. RNA, 12: 73-82

77. Huang, Y. *et al*. (2009) Structural insights into mechanisms of the small RNA methyltransferase HEN1. Nature, 461: 823-827

78. Spratt DA, Greenman J, Schaffer AG. (1999) Growth and hydrolytic enzyme production of Capnocytophaga gingivalis on different protein substrates. Oral microbiology and immunology, 14(2): 122-6.

79. Ximenez-Fyvie, L. A., Haffajee, A. D. & Socransky, S. S. (2000) Microbial composition of supra- and subgingival plaque in subjects with adult periodontitis. J Clin Periodontol 27: 722-732.

80. Fredricks, D. N., Schubert, M. M. & Myerson, D. (2005) Molecular identification of an invasive gingival bacterial community. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America 41, e1-4.

81. Teles, R. P. *et al*. (2007) A three-year prospective study of adult subjects with gingivitis II: microbiological parameters. J Clin Periodontol 34: 7-17.

82. Kira S. Makarova, Yuri I. Wolf and Eugene V. Koonin. (2013) Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Research, 41: 4360–4377

83. Anantharaman, V., *et al*. (2013) Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. Biol Direct, 8: 15

84. Kovall, R.A. and Matthews, B.W. (1999) Type II restriction endonucleases: structural, functional and evolutionary relationships. Curr. Opin. Chem. Biol., 3: 578–583.

85. Williams, R.J. (2003) Restriction endonucleases: classification, properties, and applications. Mol. Biotechnol, 23, 225–243.

86. Chakravarty, A. K., Smith, P., Jalan, R. & Shuman, S. (2014) Structure, mechanism, and specificity of a eukaryal tRNA restriction enzyme involved in self-nonself discrimination. Cell Rep., 7: 339-347

87. He,X., Ou,H.Y., *et al*. (2007) Analysis of a genomic island housing genes for DNA S-modification system in Streptomyces lividans 66 and its counterparts in other distantly related bacteria. Mol. Microbiol, 65: 1034–1048.

88. Karginov, F.V. and Hannon, G.J. (2010) The CRISPR system: small RNA-guided defense in bacteria and archaea. Mol. Cell, 37: 7–19.

89. Makarova,K.S., Haft,D.H., *et al*. (2011) Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol, 9: 467–477.

90. Hayes, F. & Van Melderen, L. (2011) Toxins-antitoxins: diversity, evolution and function. Critical reviews in biochemistry and molecular biology, 46: 386-408

91. Cook, G. M. *et al*. (2013) Ribonucleases in bacterial toxin-antitoxin systems. Biochim Biophys Acta., 1829: 523-531.

92. Chopin,M.C., Chopin,A. and Bidnenko,E. (2005) Phage abortive infection in lactococci: variations on a theme. Curr. Opin. Microbiol, 8: 473–479.

93. Dapeng Zhang, *et al*. (2012) Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. Biology Direct, 7:18

94. Yamada M, Ebina Y, Miyata T, Nakazawa T, Nakazawa A.(1982) Nucleotide sequence of the structural gene for colicin E1 and predicted structure of the protein. Proc Natl Acad Sci U S A. 79(9): 2827-31.

95. Arcus, V. L., Backbro, K., Roos, A., Daniel, E. L. & Baker, E. N. (2004) Distant structural homology leads to the functional characterization of an archaeal PIN domain as an exonuclease. J Biol Chem, 279: 16471-16478.

96. Luz-Maria Guzman, *et al*. (1995) Tight Regulation, Modulation, and High-Level Expression by Vectors Containing the Arabinose P_{BAD} Promoter. Journal of Bacteriology, p. 4121–4130

97. Rolf Lutz, Hermann Bujard. (1997) Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. Nucleic Acids Research, 25, 1203–1210

98. Danesh Moazed, *et al*. (1986) Interconversion of Active and Inactive 30 S Ribosomal Subunits is Accompanied by a Conformational Change in the Decoding Region of 16 S rRNA. J. Mol. Biol. 191: 483-493

99. Isabelle Iost, *et al*. (2001) Characterization of E. coli Ribosomal Particles Combined Analysis of Whole Proteins by Mass Spectrometry of Proteolytic Digests by Liquid Chromatography–Tandem Mass Spectrometry. Methods in Molecular Biology, 432: Organelle Proteomics

100. S.V.Kirillov, V.I.Makhno, N.N.Peshin and Yu.P.Semenkov. (1978) Separation of ribosomal subunits of Escherichia coli by Sepharose chromatography using reverse salt gradient. Nucleic Acids Research, 5(11): 4305-15

101. William M. Clemons Jr, *et al*. (2001) Crystal Structure of the 30 S Ribosomal Subunit fromThermus thermophilus: Purification, Crystallization and Structure Determination. J. Mol. Biol. 310: 827-843

102. Maria Selmer *et al*. (2006) Structure of the 70S Ribosome Complexed with mRNA and tRNA. Science, 313: 1935-42

103. Matthieu G. Gagnon *et al*. (2012) Structural Basis for the Rescue of Stalled Ribosomes: Structure of YaeJ Bound to the Ribosome. Science 335, 1370-2

104. Philip J. Kranzusch *et al*. (2013) Structure of Human cGAS Reveals a Conserved Family of Second-Messenger Enzymes in Innate Immunity. Cell Reports, 3: 1362–1368

105. Andrea Ablasser, *et al*. (2013) Cell intrinsic immunity spreads to bystander cells via the intercellular transfer of cGAMP. Nature, 000: 1-5

106. Daxing Gao, *et al*. (2013) Cyclic GMP-AMP Synthase Is an Innate Immune Sensor of HIV and Other Retroviruses Science, 341: 903-906

107. Hiroyasu Konno, Keiko Konno, Glen N. Barber. (2013) Cyclic Dinucleotides Trigger ULK1 (ATG1) Phosphorylation of STING to Prevent Sustained Innate Immune Signaling. Cell, 155: 688–698

108.Palm, N.W., and Medzhitov, R. (2009). Pattern recognition receptors and control of adaptive immunity. Immunol. Rev., 227: 221–233.

109. Elie J. Diner, et al. (2013) The Innate Immune DNA Sensor cGAS Produces a Noncanonical Cyclic Dinucleotide that Activates Human STING. Cell Reports, 3: 1355–1361

110. Andrea Ablasser, *et al*. (2013) cGAS produces a 2'-5'-linked cyclic dinucleotide second messenger that activates STING. Nature, 498: 380-384

111. Xu Zhang, *et al*. (2013) Cyclic GMP-AMP Containing Mixed Phosphodiester Linkages Is An Endogenous High-Affinity Ligand for STING. Molecular Cell, 51: 226–235,

112. Jesse Donovan, Matthew Dufner, and Alexei Korennykh. (2013) Structural basis for cytosolic double-stranded RNA surveillance by human oligoadenylate synthetase. Proc Natl Acad Sci USA, 110: 1652–1657